Testing English as a Foreign Language

Two EFL-Tests used in Germany

Sebastian Kluitmann

Philologische Fakultät

Albert-Ludwigs-Universität Freiburg

Name: Sebastian Kluitmann
Anschrift: Staufener Straße 33
79115 Freiburg im Breisgau

Erklärung zur Wissenschaftlichen Arbeit

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angabe der Quellen als Entlehnungen kenntlich gemacht worden sind.

Datum: 15.04.2008

Unterschrift:

Testing English as a Foreign Language

Two EFL-Tests used in Germany

Wissenschaftliche Arbeit
zur
Erlangung des ersten Staatsexamens
für das Lehramt an Gymnasien
der Philologischen Fakultät der
Albert-Ludwigs-Universität
Freiburg im Breisgau

vorgelegt von

Sebastian Kluitmann
geboren in
Freiburg im Breisgau

April 2008

Anglistik

Acknowledgements

This paper would not have been possible without the help of many people.
I would like to thank Prof. Dr. Bernd Kortmann for supervising this thesis. Thanks are also due to Ralf Meurer of Sprachenmarkt.de for prompting the idea for this paper as well as making part three possible. I am indebted to Dr. Glenn Fulcher, who was kind enough to share his work with me. My thanks go to Sheila Gordon-Schröder for providing valuable feedback on various parts of the manuscript. I would also like to thank Ellie Purkis, who kindly proofread the manuscript.
Last but not least I would like to thank all family and friends for general support and encouragement.

Table of Contents

Preface

In our globalised world, being able to speak one or more foreign languages is a prerequisite, as employers on a national as well as on an international scale pay attention to the foreign language skills of their future employees (cf. Morfeld 2003: 385, Sommer 2005: 3 and Bauer/Toepfer 2004: 20), focusing mostly on English. For English is still the undisputed lingua franca of the modern work force; even despite the European Council's attempts to further the diversity of languages. Actually, the European Council acknowledges that in their own institution, "there have always been limits on multilingualism [...which] are dictated by both practical considerations and budgetary constraints, in the interests of keeping operating expenditure down" (COE 2007). Needless to say that Global Players, internationally and multi-culturally organised viz. oriented companies do not think differently. This is why for "millions of learners around the world the ability to communicate in English is the passport to economic prosperity, social mobility and educational advancement" (Fulcher 2007). So, to increase their chances on the job-market, they devote both time and money to having their English language skills assessed and attested.

In the following paper, I will scrutinise and compare two tests of English as a Foreign Language: ETS' TOEIC-test on the one hand and the German Ministry of Education's KMK-Fremdsprachenzertifikat. For in the case of the TOEIC-test, there has been relatively little independent research (cf. Cunningham 2002: 1) despite its apparent popularity – in fact, with more than 4 million test-takers worldwide each year, it is the most widely used test for English as a Foreign Language. The situation for the KMK-Zertifikat is even more remarkable. Although it has been offered since 1998 and is passed off as being internationally recognised, measures to ensure the appropriateness of its design and results as well as the consistent application of standards have only recently been implemented (cf. Ó Dúill et. al. 2005: 1). Furthermore, apart from the aforementioned study, which was conducted by the developers of the test

themselves, there has been no evaluation of the KMK-Zertifikat whatsoever. The decision to choose the TOEIC-test and the KMK-Zertifikat was thus motivated by the evident lack of independent research regarding the respective tests as well as a keen interest in the comparison of perhaps differing test designs and test methods.

In the first chapter 'Language Testing', I will provide an introductory summary of the current state of research by investigating language testing in general and the intricacies and problems involved, touching issues such as the different sorts of test design/method, the phenomenon of washback, reliability and validity as well as more ethical considerations questioning the use of tests or even the very standards they are based on. Once this groundwork is laid, I will begin with the actual comparison of the two tests in chapter two 'ETS' TOEIC and the KMK-Zertifikat', matching them against the criteria developed in chapter one. Chapter three 'Survey' concludes with a survey of language testing at school and the two tests' popularity based on an empirical study conducted from September to November 2006. 230 grammar schools (Gymnasien) in Baden-Württemberg were contacted and presented with a questionnaire concerning the significance of foreign languages for the particular school, the interest in and present use of language tests as well as the popularity of major tests of English as a foreign language. The 142 (61.74 %) questionnaires that were sent back provide the foundation of the third chapter. Apart from the obvious question of how familiar the teachers are with the two tests, I will examine whether there is a correlation between certain basic background conditions such as the quantity of pupils, the location of the school or the implementation of foreign languages in the school profile and, say, the availability of external certification at this school or the teachers' degree of familiarity with different tests. In any way, the teachers can be seen as playing an important role in promoting certain tests, as pupils are likely to ask them which test they should take for their specific purposes.

However, an area which might merit further research is the question of the popularity of various different tests with employers, viz. recruiters, as existing information tends to focus either on the acceptability of tests in a particular, often academic, context[1], or is restricted to a particular test itself[2]. Thus, at the moment, a comparison of the tests in this respect is impossible. First tentative investigations in this area can be detected in the TOEIC marketing material on the one hand and Wagner's study as cited by Ó Dúill (Ó Dúill et al. 2005: 5/6) on the other. The respective findings will be covered in more detail in chapter 2.

[1] Consider e.g. universities' criteria for admission to certain programmes: some accept the TOEFL only, some demand that future students take the IELTS, others are more liberal and allow various different tests as well.

[2] Although studies may reveal that in their ads, a certain number of companies refer to a particular test's results to describe the applicants' desired English language abilities, it has yet to be shown that candidates with similar skills, yet a different certificate proving them would not be considered.

Introduction – the history of language testing

The history of testing can be traced back a long way. Play is one of the basic phenomena constituting the human condition (cf. Fink 1995: 356 ff) and in play, humans compete against each other and test their abilities. Game and play reveal the positive aspect of testing from the testee's perspective, which is often forgotten. However, testing in its broader sense is part of our everyday life. In playful activities, we learn to set and achieve goals, to enjoy victory as well as to cope with defeat.

Acounts of language testing can be found throughout the history of mankind. Probably the first evidence is found in the Old Testament, when the Gileads use a *Shibboleth* to distinguish between friends and enemies (cf. Kunnan 1999: 707, Brown / Hudson 2002: 1).

> And the Gileadites took the passages of Jordan before the Ephraimites: and it was so, that when those Ephraimites which were escaped said, Let me go over; that the men of Gilead said unto him, Art thou an Ephraimite? If he said, Nay; Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand (Judges 12: 5-6).

This story also indicates that, in testing, there is always a standard people are expected to meet. In this case, it was the standard of being able to pronounce the word "shibboleth" correctly. The existing problems concerning standards and standard-setting will be touched on in later sections.

As we have seen, the beginning of language testing dates back more than two millennia and for certain purposes, not all as sanguinary as in the biblical tale, people have always relied on certain language tests. The history of modern language testing, however, is relatively young. The growing demand for soldiers

with foreign language skills due to World War I caused the US army to develop the Army Alpha tests as a tool to measure people's linguistic abilities. Nevertheless, it was not until the 1940s, that language testing became an object for scientific research, with "Vilareal's *Test of Aural Comprehension* in 1947 and Lado's *Measurement in English as a Foreign Language* in 1949" (Kunnan 1999: 707) being the first two Ph.D. dissertations in the field of language testing. Robert Lado went on to do further research and in 1961 presented his views in *Language Testing*. His structuralist approach promoted discrete point testing, a concept which "was reinforced by theory and practice within **psychometrics** [sic]" (McNamara 2000: 14). It is based on the assumption that the four basic language skills listening, reading, writing and speaking are independent from one another and can therefore be assessed separately. In contrast to this, the notion of integrative testing acknowledges the fact that mastery of a language's formal system alone is not enough. For this reason, integrative tests aim at adding a context-specific component to the purely decontextualised discrete point testing format[3]. Typical tasks include essay writing (e.g. as a response to a given statement or essay) as well as an oral interview. These tests may thus reflect natural linguistic behaviour more accurately and feature a more appropriate theoretic approach, but, as judges are needed to assess the individual test samples[4] "such integrative tests tend[] to be expensive […] and in any case […] potentially unreliable" (McNamara 2000: 15). In the late 1970s, as an answer to these problems, John Oller proposed so-called pragmatic tests on the basis of his Unitary Competence Hypothesis. He was convinced that language proficiency

[3] Although integrative testing must be seen as a reaction to discrete point testing, it seems that in "most cases what was proposed was an adjunct to the discrete-point methods rather than their complete replacement" (Baker 1989: 65).

[4] Different theories (e.g. Rasch Measurement, Item Response Theory), which will be introduced in 1.2, have been refined in order to find a way to tackle the problem of rater-subjectivity.

was indivisible and could consequently not be tested in part. Pragmatic testing formats, such as cloze tests or dictation, related to language proficiency as a unitary concept and thus reflected language ability more aptly, he claimed. Further research revealed, however, that due to various reasons the Unitary Competence Hypothesis had to be given up, although a weaker version supporting "the 'partially divisible' nature of language proficiency" (Baker 1989: 72. Also cf. Kunnan 1999: 708, and McNamara 2000: 15) could be maintained.

Yet another concept gained popularity during the 1980s: the Communicative Testing approach was first formulated by Canale and Swain and took into account the "interactive, direct and authentic" (Shohamy 1997: 142) nature of language. With regard to actual testing, this means that we should aim at providing candidates with authentic stimuli and tasks as well as testing them while they are engaged in some sort of communication, be it active or passive. As for the theory of Communicative Testing, Canale and Swain divided general language competence into four sub-competencies: grammatical competence, strategic competence, discourse competence and sociolinguistic competence, a concept which was elaborated by Lyle Bachman in 1990 and revised by Bachman and Palmer in 1996. Their model holds that general language knowledge consists of two sub-domains, organisational knowledge including grammatical and textual knowledge on the one hand, AND pragmatic knowledge including functional and sociolinguistic knowledge on the other (cf. Bachman/Palmer 1996: 68). "Communicative language testing dominates the field" (Shohamy 1997: 143) and it is Bachman and Palmer's concept which "has been accepted as the definition of language competence used by testers that is often used as a basis for test construction"(ibid.) although it somewhat lacks empirical evidence (cf. Kunnan 1999: 707. and Shohamy 1997: 146). Portfolio evaluation, observation and simulation among other methods are used by a strand of language 'testing' called alternative assessment which holds that a person's language skills cannot be sufficiently assessed by a single test as general language ability is far too complex. Even the best test can only hint at what the testee can really do.

Therefore, "multiple assessment procedures […] are needed so that a broader and more valid language perspective can be sampled (Shohamy 1997: 142).[5]

[5] Although Shohamy claims that „performance assessment and alternative assessment are becoming more and more common" (Shohamy 1997: 143), I am aware of only two 'tests' trying to apply the principle of alternative assessment on a large scale: The Trinity College GESE and ISE certificates and the UniCert examinations.

1 Language Testing

1.1 What is a language test?

In a broad sense, a lot of tests can be classified as language tests, ranging from exams at school (e.g. vocabulary tests, grammar tests etc.) or university to certificates aiming to provide the holder with some sort of standardised qualification. In this paper, I will only be dealing with certificates which claim to provide a fair, standardised assessment as the basis for wide recognition. I will not be concerned with the testing of language within a restricted classroom environment, as clearly, this has to follow different rules, meet different necessities and is much more likely to employ alternative forms of assessment.

In general, we can distinguish two kinds of tests: proficiency tests assess the amount to which the testee has reached 'proficiency', i.e. a certain predefined level, while achievement tests usually follow the principle of test as you teach (cf. Vollmer 2003: 365). Therefore, it seems much easier to prepare for as well as to interpret the results of achievement tests. However, to be of any use for successful testees of language tests, it should be possible for, say, potential employers to draw inferences from the obtained certificates or level in a given test to the candidates' actual language skills. This is what proficiency tests do. Whereas achievement tests look backwards in that they assess what should already have been learnt, proficiency tests tend to look forward in that they assess a person's language skills and allow for interpretations of their future performance to be made[6]. This is why many, if not all, 'official' language tests or certificates are proficiency tests. Usually, the successful candidate is supplied with an interpretation grid accompanying the actual certificate to show which

[6] A problem which will be important when evaluating the two tests is the effect of washback, i.e. the influence of (usually high stake) tests on the preceding teaching, which will be explained in 1.3 and discussed later on with regard to the individual tests.

tasks typical candidates obtaining a particular level are able to fulfil. Apart from statistical errors, however, problems can also arise due to the basic framework the tests refer to and the way in which this link is established. In the case of the TOEIC-test as well as for the KMK certificate, this framework is the Common European Framework of Reference (CEF).

However, it is not enough for language tests to refer to abstracted outside descriptors of language competence, they also have to prove their reliability as well as their validity.

In language testing, reliability means that the test really functions consistently, whereas validity indicates the amount to which a testee's test result is true, i.e. whether it correctly reflects the testee's actual language ability.

1.2 Reliability and validity

In this part, I will introduce the concepts of reliability and validity as well as some of the statistical intricacies involved. Apart from discussing the various aspects of validity, their linkage to the notion of washback and impact, and the ways in which validity is linked to reliability, I will touch on the issue of reliability itself. In particular, I will outline several measurement theories and their respective advantages, viz. disadvantages (CTS [kr20, kr21], G-theory, IRT).

1.2.1 Reliability

A test is said to be reliable if it consistently yields similar, or ideally even the same results when conducted multiple times under the same conditions. Reliability $r_{TT'}$ is thus defined as the correlation between the results of one test administration T and another test administration T' under the same circumstances. The higher the correlation, the more reliable is the test. To ensure the test takes into account only systematic factors[7], that is for example the test takers skills, test developers aim at reducing measurement error, i.e. unsystematic influences on the test performance like "lapses in students' concentration, or distracting noises in the examination hall"(Alderson, Clapham, Wall 2005: 87). Normally, however, a certain degree of variation has to be expected since it is virtually impossible to rule out variation of all of the many factors involved in a test taker's performance.

[7] Test takers personal conditions can also be classified as systematic (cf. Bachman 1990: 164-166).

Classical True Score theory

This is reflected in the Classical True Score (CTS) theory's hypothesis that a testee's actual score consists of two "components: a *true score* that is due to an individual's level of ability and an *error score*, that is random and due to factors other than the ability being tested" (Bachman 1990: 167).

This yields the equation (1):

$$X = T+E,$$

with X being the actual observed score, T the true score and E the random error score. The reliability of a test $r_{TT'}$ is therefore the proportion of the observed score variance s^2_X that is true score variance s^2_T (for this and the subsequent calculations cf. Bachman 1990: 170 ff, Brown/Hudson 2002: 151 ff and for more details cf. the excellent website http://mathworld.wolfram.com), which leaves us with the following equation (2):

$$r_{TT'} = s^2_T / s^2_X$$

Considering (1), we note that the observed score variance s^2_X is the true score variance s^2_T plus the error score variance s^2_E. Thus (3):

$$s^2_X = s^2_T + s^2_E,$$

therefore (4):

$$s^2_T = s^2_X - s^2_E$$

Inserting (4) into (2) leads to the definition for reliability (5):

$$\begin{aligned} r_{TT'} &= [s^2_X - s^2_E] / s^2_X \\ &= [s^2_X / s^2_X] - [s^2_E / s^2_X] \\ &= 1- [s^2_E / s^2_X], \end{aligned}$$

solving this for s^2_E yields the definition for error score variance[8] (6):

$$s^2_E = s^2_X [1 - r_{TT'}],$$

which helps us calculate the standard error of measurement SEM (7):

$$SEM = s_X \{\sqrt{[1 - r_{TT'}]}\}.$$

The SEM enables us to make inferences about a particular candidate's true score, whereas the reliability index is relevant only for sets of scores (cf. Bachman 1990: 171). Given the Gaussian normal distribution, which in CTS is assumed as a prerequisite, there is a 68 % probability that a test taker's true score lies within the range of ± 1 SEM. The likelihood that the true score lies within the range of ± 1.96 or ± 2.58 SEM is even greater, namely 95 % and 99 % (cf. Fulcher/Davidson 2007a: 108, Bachman 1990: 197-201).

As said before, the initial definition of reliability referred to several administrations of one test under the same circumstances, also known as test-retest reliability. There are, however, some problems with this model: Firstly, it is difficult to keep the conditions similar, as this means testing the same candidates at least twice, while it is also essential that the test takers do not change in their approach to the test. Clearly, the latter is virtually impossible to achieve, as candidates are bound to react differently for various reasons (cf. Bachman 1990: 181/182, Brown/Hudson 2002: 162/163). Therefore, calculating test-retest reliability is only reasonable in those cases where we are interested in the stability of a test, e.g. if we "would like to rule out the possibility that changes in observed test scores [are] a result of increasing familiarity with the test" (Bachman 1990: 181) and can exclude significant systematic changes as to the test takers.

[8] Variance is defined as the square of the standard deviation.

1.2.1 Reliability

Apart from test-retest reliability, there are two other ways to estimate reliability: parallel-form reliability and internal consistency reliability. Parallel-form reliability is concerned with the correlation between one test version and another, parallel one. Whereas this model solves the problem of having to present the same test to the same candidates twice, it creates another: having to come up with a parallel test version of equal difficulty and standard deviation (cf. Bachman 1990: 183). Although, in the case of official language tests, there should be an item pool big enough to create multiple equally difficult test versions and there might even be the need for alternate forms, we can think of other, less official settings, in which this model is impractical. For these and many other instances, internal consistency reliability can be a solution to the difficulties both test-retest reliability and parallel form reliability pose. In internal consistency reliability estimates, one single test administration is enough to provide information about the reliability of the entire test, as the test is split in two halves which are then treated as parallel test versions. Obviously, though, we have to make sure that the two halves are equivalent in terms of difficulty, mean and standard deviation as well as independent of each other. "That is, that an individual's performance on one half does not affect how he performs on the other" (Bachman 1990: 175).
As the correlation coefficient increases with the number of individual items, the reliability index of the two halves is likely to be smaller than that of the entire test. To correct this, the Spearman-Brown formula is frequently used (8):

$$r_{kk} = kr_{hh'} / [1+ (k-1)r_{hh'}].$$

Here, k is the factor by which the length of the test is in- or decreased, r_{kk} is the reliability index of a test k times the length of the halves and $r_{hh'}$ the reliability of the halves. If we wanted to estimate the reliability of the original test, we would therefore have to put k = 2, which yields (9):

$$r_{TT'} = 2r_{hh'} / [1+r_{hh'}].$$

1.2.1 Reliability

Most of the time, however, it is hard to rule out the chance of creating two heterogeneous sets of test items. Therefore, it is more feasible to use a formula which takes into account every possible item combination such as Cronbach's alpha (10) or its more specific cases Kuder-Richardson 20 (11) and Kuder-Richardson 21 (12) and rests upon the ratio of item variance to total score variance. The most general equation is Cronbach's alpha (10):

$$\alpha = [k / (k - 1)] [1- (\{\sum s^2_i\} / s^2_X)]$$

with k being the number of items, $\sum s^2_i$ the sum of the item variances and s^2_X the total test score variance. For dichotomously scored items, Cronbach's alpha is equivalent with the Kuder-Richardson 20 formula (11):

$$r_{TT'} = [k / (k - 1)] [1- (\{\sum pq\} / s^2_X)].$$

Here, $\sum pq$ expresses the sum of the item variances, as for a dichotomously scored item, the variance is defined as the product of the proportion of correct answers p and the proportion of incorrect answers q (cf. Bachman 1990: 176). If all items are equally difficult, Kuder Richardson 21, which requires only the total score variance, the mean and the number of items, can be used to estimate reliability (12):

$$r_{TT'} = [k / (k - 1)] [1- (\{M_X (k – M_X)\} / s^2_X)].$$

In this case, M_X is the mean.

As we have seen so far, internal consistency reliability estimates can have advantages over test-retest and parallel form reliability. Nonetheless, I would not go as far as Alderson, Clapham and Wall, who almost completely dismiss the latter ones as "so time consuming and unsatisfactory" (Alderson, Clapham, Wall 2005: 88) but rather go along with Bachman's more cautious statement that the

question of which kind of reliability to estimate depends on "what we believe the sources of error are in our measures, given the particular type of test, administrative procedures, types of test takers, and the use of the test" (Bachman 1990: 184). However, internal consistency reliability estimates are much more common, since it is highly unlikely that a test should be reliable in any other respect if it is unreliable internally. Therefore, "we generally attempt to estimate the internal consistency of a test first" (Bachman 1990: 184).

That said, there are much more basic problems with all the above mentioned reliability estimates. Firstly, CTS theory reliability estimates can only ever acknowledge one possible cause for error. In other words, it "treats error variance as homogeneous in origin [... and] other potential sources either as part of that source, or as true score" (Bachman 1990: 186). Apart from that, in CTS theory, all error is supposed to be random; thus, a differentiation between systematic and unsystematic error influencing the result is impossible. The next model to be presented tries to cope with some of these shortcomings (cf. Bachman 1990: 187).

G-theory

G- or generalisability theory provides a very different approach from CTS theory, in that it does not generally presuppose a Gaussian normal distribution of test scores[9]. Rather, G-theory regards a single score as one realisation of all possible scores making up the 'universe' of scores[10]. Furthermore, G-theory can take into account many different factors influencing the actual test score, which enables us to find out whether they should be regarded as systematic or unsystematic error, or are part of the skill tested:

> The G-theory model conceptualizes a person's performance on an assessment task as a function of several different factors, or *facets*, which can include the components of language ability to be measured as well as the characteristics of the assessment procedure (Bachman 1997: 255).

An individual's ability to perform certain tasks in the real world is then estimated by drawing inferences from this individual's performance in the test, in other words, by generalising it.

On the outset, despite the differences in the concept, the generalisability coefficient $\varrho^2_{xx'}$ looks a lot like its CTS analogue, the reliability coefficient. Consider (13):

$$\varrho^2_{xx'} = s^2_p / s^2_x,$$

with s^2_p being the universe score variance, which can also be described as the person score variance, as variance due to individuals' performance is what we

---

[9] It does, however, assume a normal distribution of error (cf. Brown / Hudson 2002: 184).

[10] This aspect makes G-theory particularly valuable for developers / users of criterion referenced tests, which sometimes may not want to use the CTS model due to its assuming a normal distribution of test scores as a precondition.

aim to measure, and s^2_x being the observed score variance. Again, it is assumed that s^2_x consists of the universe score variance s^2_p, the analogue to the true score variance in CTS, and error score variance s^2_E, leading to (14):

$$\varrho^2_{xx'} = s^2_p / (s^2_p + s^2_E).$$

Here, it is noteworthy that s^2_E is usually labelled differently to distinguish between the use in a norm-referenced test as opposed to a criterion referenced test, because of the underlying assumptions in the respective concepts. Therefore, the standard error variance in NRT is often described as s^2_δ, whereas it is s^2_Δ in CRT. In contrast to the above mentioned formulas (KR20 etc.) which were only applicable in a NRT context, could only take into account one potential source for error and had to treat all other error as random, G-theory can incorporate various facets into its formula. Those facets are thought of as proportions of the error variance s^2_E. Thus, if we were interested in the influence of different forms, different raters and the effects of the different forms, viz. raters on the testees, the formula would look like this (15):

$$\varrho^2_{xx'} = s^2_p / (s^2_p + s^2_f + s^2_r + s^2_{pf} + s^2_{pr}),$$

where s^2_p is the universe score variance, s^2_f the variance accounted for by the differing forms, s^2_r the variance due to different raters, s^2_{pf} the variance that can be accounted for by the interaction of testees and forms, and s^2_{pr} the variance due to the interaction of testee and rater (cf. Bachman 1990: 192-194 and Brown/Hudson 2002: 181/182). If we were interested in finding out about the influence of more or less facets on the test scores, we could simply add them to or subtract them from our calculation.

All in all, G-theory can be a powerful and flexible means to calculate reliability estimates.

Item response theory

Item response theory (IRT), a term often used to subsume several different models, represents yet another approach to reliability which focuses on the individual item difficulty[11]. In order to meaningfully employ IRT, certain preconditions have to be fulfilled. In comparison to the other two theories just presented, these assumptions are much more specific and restrictive but, in turn, allow for much more specific inferences concerning a test taker's actual ability to be made. The first condition is that of *unidimensionality*, i.e. the assumption that each item taps one specific skill, assesses one single *latent trait*[12]. Secondly, IRT presupposes *local independence* of the test items, which means that the testee's performance on one item does not depend on his performance on another.

In terms of application, users of IRT would first estimate the items' respective difficulty, viz. their facility values. *Item difficulty* is defined as the proportion of testees correctly answering an item. Its *facility value* is given on a scale from 0 to 1 with 1 for an extremely easy item which all test takers were able to get right and 0 for an extremely difficult item which no test taker could answer correctly. Obviously, none of the extremes is of any use in classifying testees, which is why it "is generally assumed that items should not be too easy or too difficult for the population for whom the test has been designed. Items with facility values around 0.5 are therefore considered to be ideal, with an acceptable range being

---

[11] Mathematically, though representing differing approaches, the basic Rasch model and basic IRT models are identical (cf. Pollitt 1997: 244; cf footnote 13).

[12] Therefore, some authors prefer the term latent trait theory to IRT, especially when including Rasch measurement, for "this exclusive concern for *items* [in IRT] is so alien to Rasch's principle of *simultaneous* definition and measurement of ability and difficulty, or the essential symmetry of the facets, that it is inappropriate to include Rasch models under the term 'IRT'" (Pollitt 1997: 244/245). For more specific information compare Pollitt's presentation of Rasch measurement to Bachman's or Fulcher/Davidson's account of IRT.

from around 0.3 to 0.7" (Fulcher/Davidson 2007a: 102; cf. Bachman 1990: 207 and McNamara 2000: 61). Once the facility values have been established, the test items are arranged on another scale according to their difficulty. From the testees' performance on those items we can then draw conclusions about their ability, which can also be expressed as a value on the same scale. "As such, there is a direct connection between ability and difficulty" (Fulcher/Davidson 2007a: 109). This makes IRT models very convenient, as they are able to provide information on a test taker's latent trait, his ability, directly. It is also one of the aspects that positively distinguish them from CTS-and G-theory, which could only make inferences about a person's actual ability based on the performance of a sample group, whereas IRT estimates are sample independent. Apart from that, IRT models can incorporate more information than CTS- or G-theory. Usually, separate standard errors are attributed to the individual items, which again aids the interpretation of the test takers' results (cf. Fulcher/Davidson 2007a: 109). As both CTS and G-theory depend on groups, this would be impossible in either approach. Furthermore, IRT models can be selected according to the data in order to ensure the best possible model-data-fit[13]. That is to say that depending on the data, i.e. the testees' answers, the analysis is conducted using the most appropriate model. In some cases, this may be a one parameter model presupposing equality of discrimination indices for all items and ruling out the possibility of correctly answering a question by chance. In others, it could be a two- or multiple parameter IRT model taking into account more factors (cf.

[13] Again, there is a difference between proponents of 'real' IRT models and those of Rasch measurement. While those in favour of IRT accept taking into account more than one parameter, those favouring Rasch measurement rather exclude data which is not in line with the theory. While IRT proponents aim to adjust the model, Rasch proponents aim to adjust the data. "The basic issue is whether one begins by asking whether the data fit the model [...] or whether the model fits the data"(Brown/Hudson 2002: 207).

Bachman 1990: 207, Brown/Hudson 2002: 207/208). The better the model-data-fit, the more appropriate are the inferences based on the test[14].

In summary, it can be said that IRT is probably the most useful model for making inferences about a test-taker's actual language ability. However, the strong assumptions on which IRT is based can sometimes make it inappropriate to use. In some instances, it will be questionable whether the precondition of unidimensionality is fulfilled, in others, the notion of local independence may be violated. The increased value attributed to authentic test tasks triggered by the current appreciation of communicative competence and theories underscoring communicative language ability have led to test items which are mutually interdependent (cf. Bachman/Eignor 1997: 230-232). Just as in real-world tasks, these test tasks might be arranged around one central topic, e.g. a business letter. In this case, the testee might be asked to answer questions on the text or give a summary, and to follow up by composing an answer, viz. making a telephone call. Clearly, here, the test taker's performance on the latter part is not independent from her[15] performance on the first. So, despite the obvious advantages of IRT over CTS and G-theory, there are cases in which it is not feasible to conduct a reliability study based on IRT[16]. In such cases, one of the other models should be used to calculate the reliability of the test in question unless the assumptions underlying those theories are violated as well. In order to avoid some problems concerning local independence in IRT, grouping items to

---

[14] Those interested in the operation of fitting the model to the data are referred to Brown/Hudson 2002: 207-210.

[15] In an effort to keep gender neutrality, non-gender-neutral pronouns such as he/she etc are used interchangeably.

[16] Apart from the above mentioned problems, one also has to take into account the relatively high number of test-takers needed to conduct IRT studies. The figures needed to rule out statistical error are, according to Alderson, Clapham, Wall, 100 participants for the 1-parameter model, 200 for the 2-parameter model and 1000 for the 3-parameter model (cf. Alderson, Clapham, Wall 2005: 91).

form so-called testlets which are then treated as individual items can also be an option (cf. Brown/Hudson 2002: 206/207 and Bachman/Eignor. 1997: 231). However, to satisfactorily overcome the difficulties involved in estimating the reliability of performance tests, particularly those focusing on the notion of communicative language ability, it may be necessary to develop an entirely new approach (cf. Bachman/Eignor 1997: 231/232).

1.2.2 Validity

In language testing, validating a test means being able to establish a reasonable link between a test-taker's performance and her actual language ability. So, the question in validating a test is: "Does the test measure what it is intended to measure?" (Lado 1965: 30). As reliability ensures the consistency of a test, its being reliable is a precondition for its validity. For how can we learn anything about a person's language ability if the test does not even yield consistent results (cf. Alderson, Clapham, Wall 2005: 187)? In fact, talking of **a test's** validity is quite misleading since what is validated is not the test itself. Rather, it is a matter of validating the inferences we draw and "the interpretations and uses we make of test scores" (Bachman 1990: 236, cf Banerjee/Luoma 1997: 275 and Brown/Hudson 2002: 212). Validity, then, can be seen as a concept allowing us to endow test-scores with meaning. This unitary notion of validity has traditionally been subdivided according to the kind of evidence on which the interpretations are based. Usually, one will come across the terms 'construct validity', 'content validity', 'criterion-oriented validity', 'concurrent validity', 'face validity' and 'consequential validity'. It should, however, be understood "that these 'types' are in reality different 'methods' of assessing validity" and "that it is best to validate a test in as many ways as possible" (Alderson, Clapham, Wall 2005: 171).

Furthermore, one has to understand that, in interpreting test-scores, even the most valid and reliable test can only reveal what the testee is able to do, but not, what he cannot do. For even the best test cannot rule out the possibility of the test-taker's suboptimal performance due to factors unrelated to the test (cf. Bachman 1990: 146). That a testee is unable to fulfil a certain task in a testing situation does therefore not necessarily mean that he is unable to fulfil this task in real life.

Construct validity

Probably the closest to the starting question of validity, does the test measure what it is intended to measure, construct validity looks at the theory or construct the test is based on. The construct is defined as the abstracted set of abilities we want to infer from the test results. So, before asking whether the test measures what it is intended to measure, one has to be clear about 'what it is intended to measure', has to be clear about what the test construct is. Only then can we ask what the test actually measures and compare it to the predefined construct. Especially when the construct appears to be somewhat questionable, it is important to bear in mind that "the theory itself is not called into question: it is taken for granted. The issue is whether the test is a successful operationalisation of the theory" (Alderson, Clapham, Wall 2005: 183). Put simply, we must not be misled by claims of high construct validity if we are not convinced of the fundamental construct. Namely, what is done in verifying construct validity is looking for evidence that the test indeed taps those kinds of skills or abilities the construct specifies. "That is, in conducting construct validation, we are empirically testing hypothesized relationships between test scores and abilities" (Bachman 1990: 256). In addition to the empirical side of things, the awareness of the underlying construct also enables us to address it logically, to try to falsify it. And indeed, in what sounds like a reminiscence of Sir Karl Popper, the founder of a philosophical strand called Critical Rationalism[17], Bachman emphasises the importance of counterhypotheses for construct validity and goes on to cite Cronbach, who claimed that the "job of validation is not to support an

[17] Sir Karl Popper's views on the notion of falsification are set forth in his work *Logik der Forschung* (cf. Popper 1989) first published in 1934 (the English edition *The Logic of Scientific Discovery* was first published 1959). In his essay 'On Popper's Negative Methodology', Quine summarises them as Popper's "negative doctrine of evidence. Evidence does not serve to support a hypothesis, but only to refute it, when it serves at all" (Quine 1970: 218).

interpretation, but to find out what might be wrong with it" (Cronbach as cited in Bachman 1990: 257). Analogously, Fulcher/Davidson start their account of the philosophical history of validity with C. S. Peirce's epistemology and end with Dewey, who, like Popper, prefers using "the term 'warranted assertion', which he trades in for the notion of truth"(Fulcher/Davidson 2007a: 11).

On the empirical side, conducting construct validation is done by means of correlation. This can encompass correlating the test in question with an already established test based on the same construct, by correlating several parts of the tests with each other, by administering the test to several different groups and correlating the respective results, or by administering the test to the same group under different conditions; first before, and then after teaching them the relevant skills (cf the chapters on validity in Bachman 1990, Alderson, Clapham, Wall 2005, Brown/Hudson 2002). Mathematically, these correlations are computed by means of factor analysis or multitrait-multimethod analysis. However, as an investigation of these analyses is beyond the aim and scope of this paper, anyone interested is referred to Bachman's introduction (cf. Bachman 1990: 262ff).

Content validity

When dealing with content validity, we are concerned with "the systematic investigation of the degree to which the items on a test, and the resulting scores, are representative and relevant samples of whatever content or abilities the test has been designed to measure" (Brown/Hudson 2002: 213, cf. Moritoshi 2001: 9). Bachman identifies two aspects of content validity: content relevance and content coverage. In this case, content relevance does not only refer to the abilities the test aims to measure, but also to the test method, something which "is often ignored" (Bachman 1990: 244). Nevertheless, it is important to bear this in mind, since it can have significant effects on the test results. For example, if we think of a test for assessing someone's speaking skills, the results may vary greatly depending on whether the test-taker is required to talk to a machine (be it alone or

surrounded by other test-takers, as is the case for the TOEFL ibt) to another testee or examiner while rated by the examiner (as is the case for the Cambridge Main Suite exams), or to an administrator while being recorded on tape for later assessment (as is the case for the MFL A-level exams in Britain or the CNaVT exams for Dutch as a Foreign Language). In those instances where an examiner is directly involved, the attitude he displays can also affect the test-takers. Therefore, in developing a language test, all of the above possibilities and its respective advantages and disadvantages should be considered and carefully weighed up against each other.

The aspect of content coverage is concerned with how well the test tasks represent the tasks in the real world. So, in verifying content coverage, one needs to show that the test tasks are part of the real-world domain the test claims to cover. One possibility to do this is by "drawing multiple samples of tasks from the domain, to determine the extent to which different sets of tasks are equivalent, as a way of demonstrating content coverage" (Bachman 1990: 245). The problem with this approach is, however, that the boundaries of content domains in language testing are hardly ever clear-cut (cf. Bachman 1990: 245). Therefore, the process of 'proving' content validity usually involves 'experts' who should make their judgements "in some systematic way" (Alderson, Clapham, Wall 2005: 173). Unfortunately, though, it appears that more often than not,

> members [of an editing or moderating committee, i.e. so-called experts] opine on the content of items without much preparation, with no independent systematic approach which means that the group dynamics are likely to have a considerable influence on the outcome. (Alderson, Clapham, Wall 2005: 174)

Another problematic aspect of using expert judges in the verifying of content validity is the choice of the experts by the test developer; are they chosen because they are known to agree with each other, or are they appointed regardless of their opinion? For the testing agency developing the test, every additional day means

having to spend money. This is why, unlike "the researcher, who can afford to investigate the issue over a period of time, test developers need evidence of the validity of their instruments as quickly as possible" (Alderson, Clapham, Wall 2005: 175). Needless to say that this may, if only implicitly, put pressure on the experts and perhaps influence their behaviour.

Furthermore, the most confining aspect of content validity is its property of being exclusively test-based. By definition, it does not take into account the actual performance of testees on the test. Consequently content validity cannot give any information about the interpretation of test scores. In conclusion, demonstrating content validity is a necessary, but by no means sufficient step in evaluating a test (cf. Bachman 1990: 247).

Concurrent validity

Concurrent validity examines a particular group's results on a test in relation to some external criteria. Whereas Alderson, Clapham Wall consider these external criteria to comprise only measures of estimating language ability such as teachers' ratings, self-assessment, or other tests of the same ability, Bachman includes "examining differences in test performance among groups of individuals at different levels of language ability" (Bachman 1990: 248) as well. The latter traditionally refers to the relationship of native speakers to non-native speakers, i.e. how these two groups score on the test and whether the test can adequately discriminate them. The underlying assumption here is of course that native speakers are more proficient than non-native speakers. However, although this may be true in general, there is evidence that whether native speakers do better on tests assessing a specific trait is a different matter (cf. Bachman 1990: 248f, Bachman/Clark 1987: 29). Obviously, one also has to take into account that even native speakers will differ in their language proficiency, e.g. as a result of their education or social standing. In addition to that, there is little agreement about what a native speaker is, or, put differently, which variety of English, be it

regional or social, to adopt as 'the' standard (cf. Bachman/Clark 1987: 29).

Anyway, it is much more common to correlate test results to other external measures such as a different assessment of the same skill as opposed to native speaker performance on the same test. In this case, one should make sure that the external kind of assessment the test is correlated with has been shown to be valid and reliable itself. Apparently, "[a]lthough this may seem logical and obvious, in actual practice, it is not so easy to gather believable external data" (Alderson, Clapham, Wall 2005: 178). Even if such data can be found and matched against the test, it is sometimes questionable how this correlation aimed to support concurrent *validity* differs from correlation estimates calculated to support, say parallel forms *reliability*.

Predictive Validity

When the relationship between a test's results and the consecutive behaviour is studied and the precision with which the test was able to predict this behaviour investigated, we speak of examining the predictive validity. As for the actual procedure, predictive validation is different from concurrent validation solely "in that instead of collecting the external measures at the same time as the administration of the experimental test, the external measures will only be gathered some time after the test has been given" (Alderson, Clapham, Wall 2005: 180). However, in addition to being faced with the same problems as in trying to prove concurrent validity, predictive validity is subject to another problem: losing sight of the ability the test claims to measure in the first place. Therefore, estimating predictive validity "is problematic because the criterion behaviour that we want to predict is often a complex one that may depend upon a large number of factors in addition to language abilities" (Bachman 1990: 254) and we should not forget that "predictability does not constitute evidence for making inferences about abilities" (ibid.).

Face validity

Although including the concept of face validity in a chapter headed 'validity' is somewhat misleading, it is frequently done. In fact, however, face validity is not so much concerned with asking whether the interpretations of the test results are valid, but rather with whether they appear valid. Basically, what we are dealing with in face validity is not the actual validity but the face value test-takers and test users attribute to the test. When referring to a test's face validity, one therefore means the degree to which test-takers and –users believe the interpretation of the test results to be accurate. Face validity is therefore much more to do with acceptance than with validity (cf. Alderson, Clapham, Wall 2005: 173). Since this merely reflects the opinion of non-experts, and is influenced by factors other than the actual validity estimates, face validity "is frequently dismissed by testers as being unscientific and irrelevant" (Alderson, Clapham, Wall 2005: 172). Although this seems like a straightforward argumentation, the importance of face validity should not be underestimated. Apart from the pragmatic reason that a test is unlikely to be successful if it is not accepted by those taking or using the test, we also cannot expect test takers to be trying their best under these circumstances. "For these reasons, test appearance is a very important consideration in test use" (Bachman 1990: 289), even if Bachman himself treated it under the heading "Post mortem: face validity" (Bachman 1990: 285).

1.3 Washback

When talking about washback, we are dealing with the way in which tests affect the preceding teaching and learning process. On the one hand, washback can be seen as a negative factor in that it may add to the predictability of a test's outcome and in that it may lead to a restriction of the syllabus to only those criteria which are absolutely necessary to pass the test. Often, this is due to the tests' "content or format [being] based on a narrow definition of language ability [constraining] the teaching/learning context" (Taylor 2005: 154). In order to eliminate this and "reduce […] the learning of test-taking strategies for particular test methods" (Alderson, Clapham, Wall 2005: 46), it seems appropriate to vary test-content as well as test method. Thus, test developers aim at ensuring high validity, objectivity and fairness[18].

On the other hand, washback can have positive aspects, as well. It is particularly in effect-driven test development that these aspects become apparent. For even if

> some model(s) of language ability may (and, we would argue should) still shape the design of the test, […] what really determines the test tasks is the effect they will have: on student learning, on curriculum, on educational policy, and so forth. (Fulcher / Davidson 2007b: 231)

Therefore, if we know the effects of particular tests or test methods, they can be employed as a valuable tool to create the desired influence, e.g. in a school surrounding. However, what we actually do know about specific test washback, is "surprisingly little" (Alderson, Clapham, Wall 2005: 46). Interestingly, Alderson and Wall found that "this lack of evidence from classrooms is a characteristic of virtually all writings about the influence of tests on teaching"

---

[18] Fairness in testing should not only be seen as applying the same standards to each and every testee in grading. Fairness is just as well a matter of test-content, language variety used and test-method. These issues will be broached in section 1.5.

(Alderson, Wall 1993: 123) and, more often than not, those studies that do exist could neither confirm nor refute assumptions like the one "that performance assessments have better washback than multiple choice test formats or other individual item formats, such as cloze" (McNamara 2000: 74) but found "that washback is often rather unpredictable" (ibid.). In view of these facts, it is evident that despite its popularity the notion and nature of washback is evasive. Studies are sometimes contradictory and a thorough investigation taking into account the many extrinsic as well as intrinsic motivational factors in test preparation – both on part of the students and the teacher – is still a desideratum. Here, the concept of effect-driven test design might aid to further our understanding of washback, while at the same time leading to improved tests. However, without clear monitoring of the many variables in the test designing process, it will be hard to establish a link between positive or negative effects and the differences between tests.

1.4 Impact

Closely related to washback, the term impact "refers to any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole" (Wall 1997: 292). Whereas washback is primarily understood to concern the effects on teaching and learning, the term impact implies a broader concept. Not only does it refer to a classroom setting, but it also draws our attention to political and societal consequences (cf. Taylor 2005: 154).

Bearing in mind the difficulties of establishing a clear cause-effect relation regarding washback, such an enterprise seems even more daunting regarding general impact. Still governments and educational bodies among others believe in the impact and washback of tests and use them accordingly, viz. implement them into their policies (cf. Taylor 2005: 154). Not too long ago, Biesemann et al. stated:

> Vieles, was in Lehrplänen schon lange gefordert wird, gerät durch die Zertifikatsvorbereitung ins Zentrum des Fremdsprachenunterrichts […]. Die Beschäftigung mit Aufgabenbeispielen internationaler Zertifikatsprüfungen hat ***daher*** auch Eingang in die nordrhein-westfälischen Lehrpläne für Fremdsprachenunterricht in der gymnasialen Oberstufe gefunden […] (Biesemann et al. 2005: 6. emphasis added).

Shohamy recalls an incident taking place as early as 1985, when she spoke to the national inspector in charge of introducing a new EFL-test and points out the risks of over-emphasising the impact or effects of a test. Apparently, the inspector was only interested in the effect the introduction of the new test would have on the teaching of EFL in Israeli classrooms. He was so obsessed with the idea of fostering oral competency that he was prepared to jeopardise the correctness of the test results in order to accomplish his goal. The factors describing the test's quality "– reliability and validity – were irrelevant to [... him.] He was not

interested in whether the test gave more accurate results" (Shohamy 2001: xi). Clearly, this is inexcusable as it puts at risk the meaningfulness of a great many children's test results that could have the power to influence their future lives. Never should one trade in positive impact and washback for reliability or validity. Whenever we are trying to evaluate a test and its use, we should therefore try to reveal the motives behind it. Hopefully, we will then find that the Israeli official mentioned above was an exception.

1.5 Ethics and standards – test uses

Ethics in language testing is a concept which can be subdivided into two parts: ethics concerning the development of a test, i.e. test internal issues, and ethics concerning the use of a language test, i.e. test external issues. This distinction is also often referred to as the distinction between test bias and test fairness (cf. Spurling 1987: 410).

In terms of test internal ethics, ensuring reliability and validity are important goals. How could using a test which has not been proved to be reliable and valid be justified ethically? How can such a test be said to assess a standard (cf. Davidson, Turner & Huhta 1997: 303)? But even if a test meets the criteria of reliability and validity, there are more subtle ethical problems. Tests may unintentionally be biased for or against certain groups of language users, which can have to do with the topical content of the test or with the variety of English being tested, etc. Especially the latter aspect has generated increasing interest. With English becoming a global language, its "development is less and less determined by the usage of its native speakers" (Ferguson 1982: x). This development can be seen from two perspectives resulting in two different concepts. While supporters of the notion of 'International English' represent a rather "universalist view" in claiming that "there is one English which unites all those who use English" (Davies, Hamp-Lyons & Kemp 2003: 572), supporters of the notion of World Englishes claim that there is "a plurality of –lects" (ibid.), that there are now many different Englishes. This has important repercussions on language testing. Favouring the universalist view of International English implies that there is one common standard that can be tested. Maintaining the perspective of World Englishes, on the other hand implies a plurality of standards. But which standard is to be tested? Whose norms are to be imposed? Which standard is acceptable in answers? These questions have to be clarified before a test is introduced and raters have to abandon "the view that 'correct' English means British English, American English" (Davies, Hamp-Lyons & Kemp 2003: 274) only.

1.5 Ethics and standards – test uses

It is particularly in the realm of test external ethics that "[e]thical issues, at least under that name, have arrived only very recently on the agenda of language testers"(Hamp-Lyons 1997: 323). The problems here are also to do with the use and misuse of test scores. In this respect, the concepts of washback and impact are important, as "tests are not developed and used in a value-free psychometric test-tube [...but] are virtually always intended to serve the needs of an educational system or of society at large" (Bachman 1990: 279). So, washback and impact have also been connected with 'consequential validity', a term first used by Messick which expresses the idea "that the interpretability and meaningfulness of test scores may be compromised by the consequences of the introduction of a particular type of test" (McNamara 1999: 727).

Tests are used as a means to make politics, play an important "social and political role"(McNamara 1999: 728, cf. Shohamy 1999: 714, Shohamy 2001). Therefore, it is all the more distressing that test scores are frequently misused or misinterpreted. Of what use is a reliable and valid test, if those using it take the test score as an indication for the existence or lack of an ability the test does not and does not even claim to assess? How much responsibility do the test developers have "for the uses made of the test scores it generates" (Hamp-Lyons 1997: 326)? What if politicians are not interested in the test's quality but only in the societal effects it may have? Especially when we take the power of tests (cf. Shohamy 2001) into account, take into account the gate-keeping function many tests fulfil, we realise the importance of not only scrutinising the test in isolation, but also looking at the societal circumstances and how the test is used.

Therefore, we should not rely on claims and 'traditional' and 'established' uses of a test but should try to find out ourselves what the reasons for these uses and claims are, and who makes them for which motives.

1.6 Test and item types

Despite the importance currently ascribed to new models such as portfolio assessment and the acknowledgement that language ability consists neither of four clearly separate skills nor of one single general skill, most official language tests keep up the distinction between listening, reading, writing and speaking for practical reasons. Grammar and vocabulary are often assessed explicitly in separate sections accompanying the reading part and are obviously main criteria for the assessment of productive skills (cf. Rea-Dickens 1997: 91).

In the following sections I will discuss some issues in testing the individual skills in general as well as test items used to tap the respective skills.

1.6.1 Reading Comprehension

Any test of reading in a foreign language "should reflect as closely as possible the interaction that takes place between a reader and a text in the equivalent real life reading activity" (Weir 1997: 39). Therefore, it is essential that the text be authentic but does not presuppose inappropriate background knowledge or restrict the candidates' performance due to inappropriate length. Test items have to be designed in such a way as to ensure that they correctly tap reading comprehension and rule out other factors.

As for test item types, Weir identifies multiple choice questions (MCQ), short answer questions (SAQ) and cloze procedures as "the three principle methods of testing reading comprehension" (Weir 1997: 40). Cloze tests provide an efficient, reliable and easily scorable assessment. Unfortunately, though, it is highly unlikely that what cloze tests assess is really reading comprehension. In order to answer cloze tests correctly, it is generally not necessary to grasp the overall content of the text, it seems. They "produce more successful tests of syntax, lexis and comprehension at the local or sentence level, than of reading comprehension in general or of inferential or deductive abilities" (Weir 1997: 40/41). As a test of reading comprehension, cloze procedures therefore have to be dismissed.

1.6.1 Reading Comprehension

Short answer questions probably come closest to real-life tasks involving reading comprehension, e.g. if someone asks us to summarise an article we read. However, it is quite likely that SAQ test items involve skills other than reading comprehension. Poor performance on such an item does therefore not necessarily mean poor reading comprehension. It could also be attributed to poor writing skills. Consequently, the measurement of SAQ items is always somewhat "muddied" (Weir 1997:41).

For the testing of "complete linguistic comprehension", multiple choice questions are common and "well-adapted" (Lado 1965: 234/235). For this to be true, the multiple choice items must be designed with the utmost care. It is especially important to create appropriate disctractors so as to minimise the effect of the testees' solving the item by ruling out the wrong answers. The distractors should present possible responses, both in terms of content and form. "[E]vidence of candidates being able to determine answers without reading the passage" (Weir 1997: 41, citing Bernhardt) is most likely due to poorly constructed MCQ items. However, even if the items are designed carefully, "some concern that students' scores on multiple-choice tests can be improved by training in test taking techniques" (Weir 1997: 41) as well as the statement that an increase in test scores does not necessarily reflect an "increase in language ability" (ibid.) may well be justified.

Apart from the above-mentioned item types, we should also include summaries in the list of item types used for testing reading comprehension. Even more than with SAQ items, the problem with summaries is that factors other than reading ability will influence the outcome. In addition to the difficulties mentioned for the SAQ-type above, summaries require some degree of organisational talent as well as intelligence. A disadvantage is that raters may disagree as to what needs to be included in the summary and how it should be rated:

1.6.1 Reading Comprehension

> Identifying the main points in a text is itself so subjective that the examiners may not agree as to what the main points are. The problem is intensified if the marking includes some scheme where, say, main points each get two marks, and subsidiary points get one (Alderson et al. 2003: 61).

All in all, there are objections to all possible item types, but only in the case of the cloze test do they seem strong enough to dismiss the test type as not useful for the purpose of testing reading comprehension.

1.6.2 Listening Comprehension

As Buck (cf. Buck 1997: 65) states, there are various test methods to assess a candidate's listening comprehension that "can be arranged on a continuum, based on the amount of interaction, or collaboration, between the listener and the speaker: from non interactive monologue at one end to completely interactive discussion at the other" (Buck 1997: 65). For examining comprehension, however, most, if not all, language tests focus on non-interactive tasks. This is why transactional language is attributed greater importance. While "it is important to note that this emphasis often misses important aspects of successful listening" (Buck 1997: 66) it is just as important to be aware of the restrictions of assessing interactive listening. For as interactive listening can hardly be tested separately, interactive listening and interactive language tend to be tested as a part of an interview or information gap activity and may therefore be seen as falling into the realm of evaluating speaking ability.

In terms of test items, there are basically three different types: MCQ-format, SAQ-format and summaries. Both MCQ-format and SAQ-format are subject to the problems already discussed in section 1.6.1 but can be appropriate in many contexts. As for summaries, concern that factors other than listening ability influence the outcome is warranted. In the case of listening comprehension, this is even more important to note, for, in addition to factors as intelligence, concentration, etc, writing skills will surely influence the result of a summary. Whereas it has been claimed that there is a relatively strong correlation between reading and writing skills, this correlation is unlikely to occur for listening and writing skills. Therefore, when constructing listening comprehension test items, one should be wary of summary questions.

1.6.3 Writing

In terms of methods, the assessment of writing seems easy: the candidate is to produce written work. What is less clear is what kind of composition is to be asked for and how it is to be elicited. While most tests take the form of a question which the testees have to answer by writing a more or less short essay, letter etc., it is also possible to guide the candidates by providing verbal or picture cues. Normally, this input decreases with an increase in the level the students are expected to have.

Apart from these kinds of composition tasks, it is also possible to use a "series of pictures or topics in the native language of the students or in the goal language to stimulate a variety of responses instead of a single composition" (Lado 1965: 250). However, this latter kind of assessment which tests writing skills indirectly, through discrete test items, has often been criticised (cf. Cumming 1997: 51) and is not frequently used any more. For usually, what we understand by someone's writing skills is the ability to "write extended stretches of meaningful, literate discourse in the language being evaluated" (Cumming 1997: 52) and not proficiency in certain linguistic features of writing as correct grammar and vocabulary (cf. ibid.).

Nevertheless, composition tasks are subject to criticism as well:

> One question is whether such tasks can solicit sufficient indications of students' writing proficiency, [... a] second controversy is over whether those tasks correspond realistically to real-world writing demands, particularly where academic tasks might typically be done over periods of weeks or months (Cumming 1997: 56).

In those cases where we want to gain insights into a candidate's ability to compose academic texts, portfolio assessment is probably a more valid and appropriate tool. Unfortunately, though, it seems to be much less feasible.

Another main problem of tests assessing productive skills such as writing is

marking. In any large-scale test, there will be more than one rater involved and normally, without training, individual raters will differ in their scoring, even if the sample and criteria are the same. Therefore, the aim is to create and maintain high 'inter-rater reliability'[19] by ensuring not only that all raters apply the same criteria, but also that they understand these criteria in the same way.

[19] Inter-rater reliability is understood as the consistency of scores for the same sample between different raters. If a test were marked twice, by different raters, would the score remain the same? In the case of perfect inter-rater reliability, the score would not change, in the case of high inter-rater reliability differ slightly, and in the undesirable case of low inter-rater reliability, scores would differ greatly. Regarding the calculation of reliability estimates see section 1.2.1.

1.6.4 Speaking

"The ability to speak a foreign language is without doubt the most highly prized language skill and rightly so" (Lado 1965: 239) Lado once remarked. The reason for this is probably the same as for the difficulty of testing speaking: its complexity. Apart from having to fulfil all the criteria which have to be fulfilled in written texts as well, spoken language leaves little time to think, utterances are made spontaneously. Furthermore, in order to be perceived as a proficient speaker of a language, one has to get the intonation, the pronunciation and prosody right. Moreover, one should be able to adapt the register according to one's interlocutors. All of these criteria should also be considered in at test assessing speaking skills. The complexity of the notion of speaking ability as well as the need for raters, who may differ in their scoring pose some of the problems in assessing students' speaking skills. Therefore, issues "such as sources of measurement error, generalisability and score interpretation, are more critical in the testing of speaking than in any other type of language test" (Fulcher 1997: 75). As for item types, it has been found that "neither the nature nor the degree of the effect of tasks on scores from tests of speaking are well understood" (Fulcher 1997: 80), which is why it seems advisable to make use of more than one item type (cf. Ó Dúill 2006: 133, Fulcher 1997: 79). It is also highly recommended that candidates be tested individually as opposed to groups or pairs, since research "suggests that it may not be fair to assign scores to individuals in group assessment" (Alderson/Banerjee as cited in Ó Dúill 2006: 135).

Regarding the actual scoring process, what has been said about the scoring of productive skills in the section on writing skills is also true for the scoring of speech. Examiners have to be trained in such a way that they interpret and apply the rating criteria alike. In other words: high inter-rater reliability is a main concern.

1.6.5 Grammar

Due to the increasing value attributed to communicative performance, "the assessment of grammar has not been high on the language testing agenda in recent years, from either pedagogical or research perspectives" (Rea-Dickins 1997: 95). Nevertheless, grammar plays an important role in assessing the productive skills. Quite often, it is also tested in specific subtests reflecting a structuralist approach, "the best practice of the 1960s" (Rea-Dickins 1997: 93). However, since it seems to be unclear how grammar could be tested reliably and validly otherwise, it might be justified to stick to the old ways in this respect. The alternative solution is not to test grammar explicitly at all and rely on its importance for appropriate speaking and writing.

> Nonetheless, even if we knew, and we do not, that it is not necessary to test grammar as distinct from, say reading and writing, this would raise concerns about potential negative washback on teaching and a further lack of respect for the teaching of grammar (Rea-Dickins 1997: 93f).

Therefore, including test items explicitly designed to tap grammatical competence appears to be well warranted. On the theoretical side, however, there is still the problem of defining the construct of grammar. What does grammatical competence encompass and in which ways is it distinct from other skills? Where are the boundaries between, say, grammar and vocabulary? Which of the following does the concept of grammar cover? "Syntax? Morphology? Cohesion? Knowledge of the linguistic system? Language awareness? Rhetorical organisation? Ability to use syntax and lexis to express intended meanings? (Rea-Dickins 1997: 94). Recent psycholinguistic research has shown that there is no such thing as clear boundaries. Rather, all these notions are inter-related. However, much more research is needed to fully understand the nature of this inter-relatedness and to be able to design test-items accordingly.

1.7 The CEF

The Common European Framework of Reference for Languages is part of the European Council's attempt to sustain and further the diversity of languages as an important step towards mutual cultural understanding and appreciation[20]. Therefore, fostering the learning of foreign languages within the European Union is as vital as is ensuring a shared standard to facilitate the official recognition of skills regardless of national borders. While it is made clear in section 1.5 of the CEF that it, as far as language testing is concerned, aims to support the test designing process regarding only:

> the content syllabus of examinations;
> assessment criteria, in terms of positive achievement rather than negative deficiencies. (Council of Europe 2001: 6),

it appears to have been associated with much more than this since its publication in 2001. In fact, "the danger of reification is great" (Fulcher 2004) and unfortunately, it is difficult to point out the differences between various tests, when, erroneously, users "compare scores across different tests that are 'linked' to the CEF"(ibid.) relying on a common standard. Unfortunately, the idea that results on various different tests could be compared has probably been sparked off by the creators of the CEF themselves. Apart from using the CEF

> for the specification of the content of tests and examinations [and] for stating the criteria for the attainment of a learning objective, both in relation to continuous teacher-, peer- or self-assessment (Council of Europe 2001: 19),

---

[20] The CEF's authors even call on the ethnological and philosophical concept of 'otherness', a theory associated e.g. with Lévinas and a notion which, due to its complexity, I will not dwell on, here.

they go one step further when suggesting using it "for describing the levels of proficiency in existing tests and examinations thus enabling comparisons to be made across different systems of qualifications" (ibid.). In order to find out whether this is an appropriate use, we have to have another look at the proficiency levels as set forth in the CEF. We will also have to take into account the way, in which these proficiency levels were arrived at. In his article for the Guardian Educational Supplement *Are Europe's tests being built on an unsafe framework?* Glenn Fulcher directs attention to how the CEF scale was derived. He points out that teachers were presented with descriptors they had to rank for difficulty. Then, Rasch analysis was used to arrive at the individual difficulty estimates. Finally, cut scores were agreed upon so that the descriptors fit into the six CEF proficiency levels. (cf. Fulcher 2004, for more detail cf. Council of Europe 2001: 217-225). Thus, from the development of the CEF scale, it has become obvious that "what is being scaled is not necessarily learner proficiency, but teacher/ raters' perception of that proficiency" (North as cited in Fulcher 2004). There is no theoretical underpinning of the framework. Rather, the descriptors, the 'can-do-statements' were assigned to the CEF levels "on the basis of teacher judgements of perceived difficulty" (Fulcher / Davidson 2007a: 98).

In his reply to the above mentioned article, North, one of the authors of the CEF, emphasises the validity of the CEF scale while also making clear that "one should not confuse a distillation of shared subjective expertise with 'scientific truth'" (North 2004). Regarding the linking of certain measures to the CEF, he concedes:

> Of course there are different degrees of rigour in the way people relate assessments to the CEF, and it is legitimate that this should be so. One would logically expect a greater degree of rigour from an examination provider than from a language school (North 2004).

Nevertheless, theses different degrees of rigour will prove to be a problem when test users look out for a given assessment's link to the CEF but are naïve enough

not to question the establishment of this link. In fact, with the CEF levels' becoming *the* system (cf. Fulcher 2004, Fulcher/Davidson 2007b), many institutions may want to link their test to the CEF "simply to get 'recognition' within Europe" (Fulcher 2004), which entails the problem that this linking is often done intuitively (cf. Fulcher 2004). For even when we accept the CEF levels, what are we to do if a candidate fulfils the criteria for a certain level with regard to some descriptors, but fails to do so for others? In other words: "How many of the 'can dos' must we be able to do before we are in a level? That is, how can a single level summarize our performance" (Fulcher / Davidson 2007a: 100)?

For all these reasons, one has to be wary whenever encountering claims of a test's linkage to the CEF. If possible, one should critically examine how the link has been established and scrutinise whether the CEF descriptors reflect the actual test content and construct[21].

[21] In addition to the above mentioned problems, we could also conceive of a situation where a test assesses grammatical accuracy only. If a candidate achieves a certain number of points on this test, it might be linked to a certain proficiency level on the CEF, say B2. The successful candidate may then claim to have demonstrated his English language skills at level B2. The problem is, now, that people may only look at the descriptors for level B2 in the CEF reference scale and forget to examine the actual test. Thus, they can be misled to expect the candidate to have reached level B2 concerning other skills such as writing, although this skill was not assessed at all.

2 ETS' TOEIC and the KMK-Zertifikat

After providing a quick glance at the most important concepts underlying language testing in Chapter 1, I will proceed to apply them in order to evaluate two EFL-tests assessing vocational English, both of which are currently in use in Germany: ETS' Test of English for International Communication as an example for a professional body's established multiple choice test on the one hand, the German Ministry of Education's KMK-Zertifikat as an example for a fairly new test adding to the pupils' school leaving certificate. In that both certificates are specific purpose tests (often called LSP tests- Language for Specific Purposes), they restrict the topical area from which test questions and tasks are taken and thus claim to provide an evaluation of the candidates' ability to perform in similar situations in the future. Regardless of the test format, there is therefore a strong element of performance testing in both cases.

Most of the testees take the respective test voluntarily in order to pep up their CV with an objective outside evaluation of their English language skills, although some institutions require a set TOEIC-Test result for admission into certain programmes. Others accept a particular TOEIC-Test result as equivalent to various other accepted certificates such as the TOEFL or UCLES' Cambridge Main-Suite exams.

Whereas the price for TOEIC-Test administrations is determined by the local ETS office and to my knowledge ranges from 90 € to 100 € for pupils, costs for the KMK-Zertifikat vary from state to state. While it is free in many areas, some charge from 30 €up to 65 €.

2.1 The TOEIC-Test

2.1.1 The TOEIC – history and (cl)aims

The Test Of English for International Communication was developed by Educational Testing Service, a non-profit, non-government organisation which was founded in 1947 merging the three formerly mutually independent institutions the American Council on Education, the College Board, and the Carnegie Foundation. Currently, ETS operates in 181 countries and, with tests as the TOEFL, SAT and GMAT among others, has become the largest private educational testing organisation.

In 1979, ETS developed the TOEIC upon request of the Japanese Ministry of International Trade and Industry. According to ETS, the TOEIC test "measures the English communication skills of people working in an international environment" (ETS. *Technical Manual*: 1.1), providing "rapid, affordable, and convenient service, as well as [...] consistency of measurement worldwide" (ETS. *User Guide*: 8). Furthermore, the TOEIC test is advertised as being recognised as "a worldwide standard for English proficiency" (ETS. *User Guide*: 4) used by organisations, companies, language schools and institutions of higher education. As for appropriate test uses, ETS name the recruiting, promoting and evaluation of staff as well as placement of students and the monitoring of student progress among others.

Critical reception

While other tests developed by ETS, such as the TOEFL, have received a lot of scientific interest, research on the TOEIC test is scarcely found. Except for Gilfert's review in the *Internet Teaching English as a Second Language Journal (ITESLJ)*, Hirai's pioneering study for Hitachi and the valuable work at the University of Birmingham with Moritoshi and Cunningham addressing important issues regarding the TOEIC test in their respective articles, there is virtually no independent research on the TOEIC test. Both, Woodford's initial validity study and the various studies conducted by Wilson seem to be ETS-funded. The first study, the initial validity study, was conducted by Protase E. Woodford in 1979, presented to the English Speaking Union Conference on English in International Communication in 1980 and published in 1982. In it, Woodford not only estimated the TOEIC test's reliability and concurrent validity by matching TOEIC part scores against other measures of the same ability (cf. Woodford 1982), but also found a high correlation, namely 0.83, between "the TOEIC listening part score and the direct Language Proficiency Interview" (Woodford 1982: 71). The correlation between the TOEIC reading part score and the direct writing measures was equally high. Thus, Woodford concluded that, as far as speaking skills are concerned, the TOEIC part score apparently "is a good predictor of the candidates' abilities to speak English even though the objective measure tests a receptive oral skill while the direct speaking measure tests a productive oral skill" (Woodford 1982: 71). As for the writing skills, it was claimed that "a separate part score for writing is not necessary" (Woodford 1982: 71). Although to my knowledge the statements concerning reliability have never been doubted, the claims regarding the productive skills are highly disputed, as we will see later on.

In 1989, Kenneth M. Wilson's report *Enhancing the Interpretation of a Norm-Referenced Second-Language Test Through Criterion Referencing: A Research Assessment of Experience in the TOEIC Testing Context* was published. It was based

on data collected from 1979-1988 (Wilson 1989: 22). In the carefully designed study, Wilson found a reasonably high correlation between the TOEIC listening comprehension score and the LPI. The correlation was calculated for three subsamples according to the nationality of the testees and averaged .74 (cf. Wilson 1989: 40). Therefore, he concluded that

> Knowledge of TOEIC/LPI relationships represents a clear interpretive advance because it permits test users to make statistically valid inferences from employees' TOEIC scores about their levels of developed oral English proficiency [...]. It also provides better-informed perspective regarding the level and range of oral English proficiency that academically trained ESL users/learners in the TOEIC testing context can be expected to exhibit (Wilson 1989: 58).

It is particularly noteworthy, that he talks of **statistically valid inferences**, because this fact is often neglected. Although a correlation may be high, inferences are only true for a given group of people and may differ greatly for individuals. Therefore, for high stakes decisions involving testing of English as a foreign language, the correlation of .74 between the TOEIC-test and direct measures of speaking skills appears to be too low. Recruiters' may be misled and could over- or underestimate a candidate's speaking abilities due to her performance on the TOEIC-test. Wilson, however, would probably disagree, claiming in his study *An Exploratory Dimensionality Assessment of the TOEIC Test* from September 2000 that "research [...] has confirmed and extended Woodford's finding of strong and theoretically consistent patterns of correlation between TOEIC test scores and LPI ratings" (Wilson 2000: 4). In this study, he evaluates the dimensionality of the TOEIC test, that is, whether the items used really assess the skills they are said to test. In his introductory remarks, he comments on the two separate sections of the test, reflecting the division into the Listening Comprehension and the Reading Comprehension part. Wilson states that

> it has been tacitly assumed as a working proposition that the four different types of questions included in the Listening Comprehension section constitute primarily somewhat different methods of measuring the same general underlying proficiency dimension, and that this also holds for the three item types in the Reading Comprehension section (Wilson 2000: 4).

As will become clear when discussing the individual items in the sections to come, this assumption is blatantly wrong. In the aforementioned report, Wilson comes to the same conclusion. In the study, he correlated the individual item scores and found out that whereas the correlations between the Listening Comprehension items were consistently high, this was completely different for the items supposedly testing Reading Comprehension. Two of the three item types seem to be concerned not so much with reading – in fact, the subheadings are *Error Recognition* and *Incomplete Sentences* – but "appear to call for relatively specific prior knowledge of formal aspects of the English language, per se" (Wilson 2000: 23). This claim is supported by a factor analysis which shows "that a subscore reflecting performance on Reading items [...] tends to tap aspects of proficiency that are factorially independent of those tapped by a subscore reflecting performance on Incomplete Sentences and Error Recognition items" (Wilson 2000: 19)[22]. Four years earlier, Gilfert has already found the same in her article *A Review of TOEIC* for the Internet TESL Journal and states: "In the reading comprehension subtest, two subsections evaluate the testee's ability to use English grammar in a relatively formal manner" (Gilfert 1996). Further on in the review, she feels the need to explicitly point out that the TOEIC test assesses the receptive, but not the productive skills, as many test takers and test users believe in the putatively strong correlation between TOEIC test results and the testee's productive skills. However, some "examinees become experts in taking language tests, but do not learn how to use the language" (Gilfert 1996). Although Gilfert

[22] Despite these findings, the structures and headings of the TOEIC test have only recently been changed marginally.

considers the TOEIC test reliable and valid as a measure of the passive skills, she draws attention to the "artificial reality" (Gilfert 1996), in which the TOEIC test operates: "Examinees who score well on these [TOEIC and TOEFL. my expl.] tests may have self-confidence in the language classroom, but using their language skills in the real world may be quite a different thing" (Gilfert 1996).

To get further insights in the actual active abilities of TOEIC testees and the question whether inferences about productive skills based on TOEIC test results are justified[23], Cunningham created an entirely new test, the *Test of Interactive Communication*, which contains two sections assessing Listening Comprehension and Reading Comprehension, always in connection with tasks involving writing and often summary skills. Thus, it is designed to test directly the testees' interactive communicative competence. This is what the TOEIC test is claimed to test indirectly. These claims go back to Woodford's initial validity study. However, as Cunningham points out, the items used by Woodford to measure writing skills have had little to do with proper writing: the first and the last of the three subtests do not include writing as a separate ability but deal with formal grammatical issues. The first task, called 'dehydrated sentences', contains "sentence elements from which the examinee was to produce a coherent English sentence making any necessary changes or additions" (Woodford 1982: 68f), in the last exercise, the test taker is asked to translate 10 sentences into English. Only the second item has to do with writing at all. Unfortunately, though, all that is asked for is a 25-40 word business letter. Considering that this is basically all that claims of the TOEIC as an indirect measure of writing skills can draw on, that these claims are justified is highly doubtable (cf. Cunningham 2002: 16). Not wanting to dwell on the exact nature and possible problems Cunningham's study, it may suffice to recall that her findings suggest "that the TOEIC does not measure communicative competence" (Cunningham 2002: 57). Furthermore, it

[23] Cunningham is primarily concerned with the investigation of "the relationship between TOEIC test score gains and increased communicative competence" (Cunningham 2002: 1).

seems that a better TOEIC test score does not necessarily reflect better communicative competence, for "there is no positive correlation between TOEIC score gains and increased communicative competence, as measured by TIC" (ibid.).

In 2002, Hirai conducted a study comparing results of the TOEIC test with a company-internal interview at the Hitachi Institute of Foreign Languages in order to measure the correlation between TOEIC test scores and direct speaking tests. He also compared TOEIC test results with the results of the BULATS writing test. Whereas he found that over "a very wide range of TOEIC scores" (Hirai 2002: 8) Woodford's results as to the correlation between the TOEIC test and a direct speaking test could be supported, with the correlation coefficient between the interview administered at the Hitachi Institute of Foreign Languages and the TOEIC being .78, the case is totally different for smaller subsets. When, for example, looking at the subset of all students enrolled in an intermediate course taking the tests, the coefficient dropped to a low .49 (cf. Hirai 2002: 8). This is due to the many TOEIC test takers who have good receptive skills but show a great lack of "experience speaking English" (ibid.). The correlation between the BULATS writing test and the TOEIC score turned out to be significantly lower than the official ETS findings, namely .66. As was the case with the comparison of TIC and TOEIC, the low correlation between the TOEIC and the BULATS writing test can be attributed to the dramatic differences regarding the item types. Again, it has to be said that the item types used in the initial validity study are mostly inadequate for measuring someone's writing skills. Therefore, Hirai "suggests that TOEIC scores be interpreted cautiously in the organization's business context" (ibid.).

Moritoshi's paper, which probably sparked off Cunningham's study, is one of the few to broach the issue of the TOEIC test's reliability and validity not only practically – criticising the measures taken – but also theoretically. While he does not seem to have concerns about the reliability of the test concluding that "the

test-retest[24] reliability for the TOEIC® test appears to be acceptably high"(Moritoshi 2001: 14, cf. Woodford 1982: 66), he has serious reservations regarding its validity. As he points out, the underlying construct has not been sufficiently defined; actually, "no explicit operational definitions for these [i.e. reading, listening, speaking and writing. my comment] abilities, or for 'general proficiency' have been provided"(Moritoshi 2001: 9)[25]. Therefore, if it is unclear what the very nature of the test's construct is, it is impossible to estimate its construct validity. As for the TOEIC test, this is clearly the case. "The overall impression is given that the test's managers have tended to skirt around the issue of construct description, so weakening the test's construct validity"(ibid.), Moritoshi concludes. In terms of concurrent validity, he mentions another facet adding to the problems already revealed by others: Of the four tests the TOEIC was measured against, three were themselves "unvalidated and all were scored subjectively"(Moritoshi 2001: 10). Furthermore, it has not been demonstrated that the test items for, say, reading comprehension assess reading comprehension only; "no 'negative evidence' is offered to show that the test is *not* testing other, unrelated abilities" (ibid.). My personal view is that, had the test's managers sought to provide negative evidence, they would probably have had to exclude the item types 'error recognition' and 'incomplete sentences' from the section on 'reading comprehension' and in fact, in the new version of the TOEIC test, the item type 'error recognition' is left out (ETS 2006).

As to content validity, a customer needs analysis was conducted to identify "those features of the general 'target language use domain'" (Moritoshi 2001: 12, cf. *Technical Manual*: III-6) which were to be included in the test. Thus, the test developers could make sure that the test content is relevant for the test users.

---

[24] To my knowledge, though, what was calculated was an internal consistency reliability estimate, which is slightly different from a test-retest reliability estimate (cf. section 1.2.1).

[25] This might be one reason why many formal aspects of language have been included in the 'reading comprehension' part as shown earlier.

Therefore, "though it is difficult to evaluate the test's content validity from the theoretical perspective, it does appear high from the practical standpoint" (Moritoshi 2001: 12). With regard to face validity, Moritoshi notes: "The TOEIC® test probably has high face validity as a measure of listening and reading skills because the test's tasks utilise these abilities directly and overtly" (Moritoshi 2001: 13). This is also supported by the high number of test takers.

All in all, Moritoshi concludes that the TOEIC test is necessary, "both in principle and practice, as it serves the needs of employers and employees" (Moritoshi 2001:16) and proved high content and face validity. Reliability and fairness are also perceived as high. On the downside, however, the "lack of operational definitions and 'negative evidence' largely invalidates the test manager's claim of high construct and concurrent validities" (ibid.). Moreover, any claims concerning the TOEIC test as an indirect but appropriate measure of speaking and writing skills "have little proven basis and are highly dubious" (Moritoshi 2001: 17).

One of the most current papers on the TOEIC test has to do with the CEF's being attributed increasingly more value. Therefore, ETS felt the need to conduct a study *Mapping English Language Proficiency Test Scores Onto the Common European Framework*. Apart from the TOEIC test, researchers were interested in the other tests provided by ETS, namely the TSE, TWE and TOEFL. The procedure for linking the tests to the CEF was in accordance with the preliminary pilot version of the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF)* issued by the COE in 2003. The actual setting of the cut scores was done by expert panellists. They were first trained to get a deeper understanding of the CEF proficiency bands B1 and C1 they would later be asked to agree on cut off scores for. Then, as for the TOEIC test, each item was considered in terms of its difficulty, i.e. panellists estimated the probability with which a candidate at CEF level B1 could answer this item correctly. This process was repeated for CEF level C1. Afterwards, panellists had the chance to discuss and adjust their decisions. Finally, the

recommended cut off scores for the CEF levels B1 and C1, i.e. "the minimum scores necessary to qualify for the B1 and C1 levels on the CEF" (Tannenbaum/Wiley 2005: 15) were recorded. According to this study, "TOEIC B1 and C1 scaled cut scores are 550 and 880, respectively" (ibid.).

2.1.2 Description and evaluation of the TOEIC-test's design

Format

The TOEIC-test in the version evaluated[26] exclusively tests the passive or receptive skills[27]. It consists of two separate sections, each containing 100 questions in MCQ format. The first section is a Listening Comprehension and lasts approximately 45 minutes. It is subdivided into four parts: in part one, the candidates are provided with picture cues and have to find out which of the four spoken statements best matches the picture cue. In the second part, there is no picture cue. Instead, there is one spoken statement and four spoken responses. Parts three and four deal with short conversations. Part three has one question for each conversation. The difficulty is increased by asking two or more questions for each conversation in part four.

The second section lasting 75 minutes is described by ETS as "testing how well you understand written English" (ETS. *Examinee Handbook*: 4). This is slightly misleading, as one would expect a reading comprehension, when, in fact, only one of its three subdivisions is concerned with the testing of reading comprehension. The other two parts are concerned with more formal issues including mainly grammar and vocabulary.

All tasks are preceded by introductory remarks and one example.

[26] All items are taken from the TOEIC Sample Test, TOEIC Form ST-00 (see appendix).

[27] During the time this paper was being written, modules for assessing speaking and writing skills resembling those of the TOEFL ibt were being introduced. Nevertheless, the basic TOEIC-test has been changed only slightly and the speaking and writing modules are optional.

Development

In order to provide a high-quality test, ETS effected a series of measures assessing the reliability and validity of the TOEIC test.

Concerning validity, what was measured was content validity and concurrent validity. Content validity was ensured by conducting needs analyses asking "many international companies [...] about the English language skills needed by employees who use English for international communication" (*Technical Manual*: III-6). Thus, the test developers made sure that the test content was relevant for the potential test users.

Regarding concurrent validity, the results on the TOEIC test were correlated "with other established methods that purport to measure the same construct" (*Technical Manual*: III-1). However, as Moritoshi has already pointed out, the authors of the *Technical Manual* fail to properly define the construct for the TOEIC test. They probably assume that everyone has the same concept of listening comprehension and reading comprehension. That this is in fact not the case, that different test developers may have different notions of listening and reading comprehension and consequently may choose different item types assessing slightly different constructs significantly restricts any claims of the TOEIC test's construct validity. Understanding this is also vital for interpreting the value of concurrent validity estimates. These were derived by relating the TOEIC to several other tests assessing speaking, writing, listening and reading. As for speaking, apart from the study investigating the relationship between scores on the TOEIC and the LPI which yielded a correlation of .74 (Woodford 1982), results from three other tests were compared with those from the TOEIC test. The correlation between the SPEAK test and the TOEIC test was found to be .78, the correlation between the Australian Second Language Proficiency Rating (ASLPR) and the Listening Comprehension section of the TOEIC test .70 and the correlation between the John Test Part II and the TOEIC Listening Comprehension section .69 (*Technical Manual*: III-1f). For the latter two, ETS

concede that the correlation between these measures and TOEIC Listening Comprehension part was "significantly stronger than that between [...them] and TOEIC Reading Comprehension" (*Technical Manual*: III-2). Therefore, we have to conclude that, although it is not stated in the *Technical Manual*, the correlation between the TOEIC total score and direct measures of speaking abilities is significantly lower than those cited above.

In terms of studying the relationship between the TOEIC test and direct writing measures, the *Technical Manual* only mentions the test carried out in the initial validity study. This test, however, has to be dismissed as inappropriate for assessing writing skills for the reasons already outlined in section 2.1.1: it is highly doubtable that what it actually tests is writing ability. Therefore, the high correlation of .83 must be considered virtually meaningless.

Concerning the Listening Comprehension section, the test it is correlated with in the initial validity study is, again, subject to severe criticism. Firstly, it does not seem to be validated itself (cf. Moritoshi 2001, Woodford 1982). Apart from that, as Cunningham points out, although the

> input for the validity tests for reading and listening was in English, the response tasks were conducted in Japanese making interpretation of the results [...] circumspect. They may indicate comprehension abilities; they do not indicate communicative abilities (Cunningham 2002: 15).

However, if one neglects the claim of the TOEIC test's measuring communicative competence, the high correlation of .90 may still suggest that the TOEIC Listening Comprehension section functions properly. This is supported by additional evidence collected from 1996 to 1999. The comparison of results on various tests of Listening Comprehension revealed the following results:

2.1.2 Description and evaluation of the TOEIC test's design

Test	Correlation*	Sample	Year
ASLPR, Listening	.73	38 ESL students in Australia	1999
TOEFL Listening Comprehension	.84	116 business school students in France	1996
TOEFL Listening Comprehension	.88	103 ESL students in Canada and USA	1999
In-house Listening placement test	.92	26 ESL students in USA	1998
CASAS Listening Comprehension	.85	31 students in California Community Colleges	1998
Michigan Listening Comprehension Test	.76	185 ESL students in Canada and USA	1999
Canadian Language Benchmarks Assessment, Listening	.67	30 ESL students in Canada	1998

* All correlations are significant at the 0.01 level (two-tailed).

(*Technical Manual*: III-3)

Regarding the validity of the Reading Comprehension section, Moritoshi's remark that the validity test is itself unvalidated holds, as does Cunningham's criticism of its non-communicative nature. Still, if we again disregard any claims of assessing communicative competence, the correlation of .79 obtained from this test might be indicative of this section's tapping Reading Comprehension skills. ETS published additional data to support this claim in a table similar to the one above:

Test	Correlation*	Sample	Year
CASAS Reading Comprehension	.73	111 students in California Community Colleges	1998
TOEFL Reading Comprehension	.76	116 business school students in France	1996
TOEFL Reading Comprehension	.83	103 ESL students in Canada and USA	1999
CLBA Reading Comprehension	.87	120 ESL students in Canada	1998

* All correlations are significant at the 0.01 level (two-tailed).

(*Technical Manual*: III-4)

Reconsidering all of the abovementioned evidence, we can reasonably conclude that the TOEIC test seems to be an accurate measure of Listening and Reading Comprehension. In view of independent research, such as the studies by Hirai, Moritoshi and Cunningham, claims asserting that the TOEIC test would indirectly but accurately assess the productive skills speaking and writing as

well, appear untenable. Interestingly, despite the assertions that "TOEIC scores can indicate whether an employee will be able to work and interact successfully if posted to an English-speaking country" (The Chauncey Group International Ltd. 1999: 7), a study examining the TOEIC's predictive validity is still a desideratum.

The reliability of the TOEIC test was estimated in terms of internal consistency reliability and was calculated using the Kuder-Richardson 20 formula (cf. section 1.2.1). This resulted in an internal consistency estimate of .95 with a Standard Error of Measurement of 25, which indicates that the TOEIC test is highly reliable (cf. *Technical Manual*: IV-1, IV-5). According to Cunningham, the fact that "99 % confidence in the score requires minimum gains of 63.5 points [...is] weakening the test's reliability for individual scores" (Cunningham 2002: 14). This claim, however, is surprising, since she did not seem to criticise the Standard Error of Measurement (SEM) which is important if we want to look at candidates' performance at an individual level. Therefore, Cunningham's discontent with the fact that "99 % confidence in the score requires minimum gains of 63.5 points" (ibid.) is somewhat peculiar. For, as explained in section 1.2.1, this has to do with the Gaussian normal distribution and is true for all norm referenced tests: the 68 % confidence interval is 1 SEM, the 95 % confidence interval 1.96 SEM and the 99 % interval 2.58 SEM. The 63.5 points simply reflect 2.58 times the SEM of 25[28]. Therefore, the TOEIC test's reliability should not be questioned based on the data mentioned by Cunningham. If the confidence intervals appear too big, it may alternatively be criticised that the SEM is too high. However, on a scale of 0-990, as is the case for the TOEIC, a SEM of 25 seems quite acceptable to me.

[28] Obviously, this is wrong. 25 times 2.58 equals 64.5, not 63.5. However, due to reasons unknown to me, ETS seem to assume that a 99% confidence interval is already reached at 2.54 SEM (cf. *Technical Manual*: IV-5).

2.1.3 Description and evaluation of the TOEIC-test's item types

Reading Comprehension

The Reading Comprehension section is subdivided into three parts. Directions and one sample question is given for each part.

In the first part, the candidate is to complete incomplete sentences by choosing one of the four answers presented. The distractors are designed sensibly in that they present possible answers conveying meanings similar to the correct answer. It is doubtable, however, that what this item type tests is really 'reading comprehension' as most people understand it. Looking up the entry for comprehension in the Collins English Dictionary yields the following result: "**3** *Education*. an exercise consisting of a previously unseen passage of text with related questions, designed to test a student's understanding esp. of a foreign language" (*Collins*: 329) In this part of the TOEIC test's Reading Comprehension section, however, there is no passage of text to which the questions could relate. What is tested in this section is probably grammar, but not reading comprehension. Item 101, for example, focuses on the correct use of pronouns, item 103 on how to form reduced relative clauses, etc. Therefore, this section cannot be considered an appropriate measure of reading comprehension. It may, however, be useful for assessing grammar.

Part two of the Reading Comprehension section deals with error recognition. The candidates are asked to identify which of the four underlined words or phrases is incorrect. Again, there is no passage of text to which the items relate, and in some cases, it may arguably be possible to answer the questions correctly without even having to understand the meaning. In item 141, recognising the plural marking on 'hotels' is enough to choose the correct answer. It is immediately obvious that the verb form 'provides' is incorrect since there is no noun to which it could relate. In order to answer this item correctly, only rudimentary reading comprehension skills are necessary. The testee does not have to know the meaning of all parts of the sentence, nor understand the overall meaning to get

this item right. What he does need, however, is knowledge of grammar. Therefore, the items in this section are not suitable as a measure of someone's reading comprehension skills. As the items in the preceding part, they may, however, be appropriate for testing grammar.

The proper reading comprehension items make up the third part of the TOEIC test's Reading Comprehension section. They are comparatively short and contain approximately 100-200 words. In terms of content, they deal with business letters, formal invitations and the like. This reflects the needs of an office environment where the staff is required to quickly read through and understand business related correspondence but rarely has to cope with longer texts or unfamiliar subject matters.

As always, the questions are designed in MCQ-format; the distractors are chosen sensibly. The questions aim at both global and local understanding. Looking at the specific test items, global understanding is assessed in questions such as in number 164: "What is the purpose of the letter?", local understanding in others, e.g. question 166: "What is a stated advantage of the seminar?" In all cases, the distractors present possible answers drawing on related content and vocabulary.

In summary, the reading comprehension items are constructed well and can reasonably be called authentic.

Listening Comprehension

As already mentioned, the LC-section is divided into 4 parts.

For all parts, individual directions and one sample question are given. The speakers speak naturally, and predominantly feature General American pronunciation. Each item is played only once.

The first part contains a picture cue. The candidates are to listen to the 4 spoken statements and choose the one which fits best. In the case of question 1, the picture shows 2 cyclists riding next to each other on a road in the mountains. All of the distractors use vocabulary which is also represented in the picture and/or is

related in meaning. Therefore, the item can be considered well designed.

In the second part, there is neither a picture cue to accompany the sound sample, nor are the answers printed. Instead, both the speech sample and the 3 possible answers are spoken. Let us look at item no. 25. This item is not as well designed as the one discussed before. First of all, there are only three answers, which increases the chances of guessing. Furthermore, both distractors focus on the same word and answer C is probably too easy to rule out. Therefore, this item cannot be considered well constructed.

Part three features short conversations. There is one question per conversation with four possible answers. Questions as well as answers are printed in the test book. The distractors used are chosen well. They are related to the content of the conversation and thus present possible solutions. All in all, this item type can be considered an appropriate measure of listening comprehension.

Part four is similar to part three. However, instead of having to answer only one question per talk, candidates are required to answer at least two questions per item. Again, the distractors are chosen sensibly and correspond to the topic of the talk.

In conclusion, we can say that, apart from items such as item 25 in part 2, the Listening Comprehension section is designed well. The various parts use different input and vary with regard to their difficulty.

2.2 The KMK-Zertifikat

2.2.1 The KMK-Zertifikat - history and (cl)aims

The starting point for the KMK-Zertifikat was the *Rahmenvereinbarung über die Zertifizierung von Fremdsprachenkenntnissen in der beruflichen Bildung* in 1998. In it, the 16 German states agreed upon a common framework for the KMK-Zertifikat. The procedure of constructing test items and the administrative process is handled by the individual states and may differ. Interestingly, it is not a condition that the test be organised centrally – not even for one state. It is therefore possible that a teacher develops the test items for the same pupils he teaches and whose papers he later on corrects. Also, tests may vary significantly from school to school and from state to state. At the early stage of the KMK-Zertifikat, there were virtually no measures to ensure common standards. This potential problem was realised and tackled five years after the first tests had been issued. The EU-KonZert-study took place from 2003 to 2006 and sought to establish common standards regarding the actual test items (2003/2004), the assessment of student performance (2004/2005) and the actual test administration (2005/2006) (cf. Ó Dúill et. al. 2005).

According to the state of Baden-Württemberg's[29] official website concerning the KMK-Zertifikat, it is

> a voluntary exam on the basis of the KMK's master agreement from 1998 regarding the *certification of foreign language skills within vocational education* in line with the COE's 1996 initiative *Common European Framework of Reference for Language Learning and Teaching*. It can optionally be administered by vocational schools and colleges. (LIS 2006. my translation)

[29] I am referring to Baden-Württemberg, as all items in the analysis are either taken from exams conducted in Baden-Württemberg in 2005 and 2006 or best practice examples from the BLK study [eu]KonZert. All items are included in the appendix.

Furthermore, the officials promote the certificate as being recognised in 41 countries of the COE and even the government of the state of Baden-Württemberg speaks of it as an "Europazertifikat" (Landtag 2005: 10) again emphasising international recognition. A school offering the test officially advertises it as "europaweit vergleichbar" (Rieber et al. 2007: 4). The KMK-Zertifikat aims at increasing pupils' chances on the job-market. This is one point, where we can discern a basic vagueness in the KMK-Zertifikat's description. On the one hand, the aim for an independent certificate which provides successful candidates with improved prospects on the job market has us, as well as potential test users, i.e. employers, believe that it is a proficiency test. On the other hand, however, this is somewhat renounced by the statement that the KMK-Zertifikat is, by definition, "ein rein schulisches Zertifikat" (Janssen 2001a: 4) which is constructed "in Anlehnung an die Lehrpläne an der Berufsschule" (ibid.). Usually, such a test would be considered an achievement test rather than a proficiency test. Furthermore, some schools apparently devote a significant amount of time exclusively to test preparation. The KS Rottweil, for example, maintains that a total of up to 80 lessons (2/week) serves "zur Vorbereitung auf das KMK-Fremdsprachenzertifikat" (Rieber et al. 2007: 4). Therefore, the KMK-Zertifikat cannot be classified as the proficiency test as which it is passed off. At best, it can be seen as a mixture of achievement and proficiency test elements.

2.2.2 Description and evaluation of the KMK-Zertifikat's design

Format

The KMK-Zertifikat is offered at three levels corresponding to the CEF levels A2, B1 and B2. Each level is further subdivided according to the pupils' vocation. The degree of specification depends on the individual state and can go so far as to developing different tests for each occupation. The weighting of the individual test parts receptive skills, productive skills and mediation is set in the common framework as being 40% for receptive skills, 30% for productive skills and 30% for mediation, but may change from state to state within the range of 10%. In Baden-Württemberg, the weighting is 30% for receptive skills, 40% for productive skills and 30% for mediation. To pass the test, the candidate has to achieve more than 50% on both the written part as well as the oral part. The time allotted for the completion of the tasks depends on the level. At level A2, the written test lasts 60, the oral test 10 minutes, at B1, the written test lasts 90, the oral test 15 minutes and at level B2, the written test lasts 120 and the oral test 20 minutes. The oral test can be administered to individuals or groups, although in Baden-Württemberg, the administration to an individual is an exception (cf. Janssen 2001).
At all levels, candidates are allowed to use a bilingual dictionary.

Development

Apart from the *Rahmenvereinbarung über die Zertifizierung von Fremdsprachenkenntnissen in der beruflichen Bildung* itself and its echoes in Janssen and Ó Dúill et al., little information can be found about the development of the KMK-Zertifikat. It is unclear who is in charge of constructing the actual test. In those states where designing the test is up to the individual schools, it is probably the teachers who decide on the test items. In Baden-Württemberg, however, the test is issued centrally, apart from the speaking test which is designed by the individual school. The question of who created it, remains unanswered. Calling

the authority in charge resulted in their claim that the test and test items were created by "experts". As for ensuring common standards, the individual states are to 'take appropriate measures'. Again, no documentation of these measures is found. Likewise, there has not been any study evaluating the validity or reliability of the KMK-Zertifikat. This is quite ironic, for in spite of Ó Dúill's emphasising the importance of issues as test construct, reliability and validity in the introduction of the BLK study he supported academically, the study itself defines neither a test construct, nor does it mention validity or reliability estimates. All that has been done during the three years this project lasted was to agree on common standards regarding the types of tasks resulting in 'best practice' sample items and their rating criteria. How their implementation is ensured and whether their implementation is ensured at all, is not specified. Apart from its being deeply disturbing that obviously, during the first years during which the KMK-Zertifikat was offered, no effort was made to ensure common standards, it seems questionable that these standards, now they are formulated, will indeed be implemented in common.

Furthermore, Ó Dúill's academic support seems to have had little effect on what actually happened in the study. Regarding oral examinations, for example, he mentions research indicating that "it may not be fair to assign scores to individuals in group assessment" (Alderson/Banerjee as cited in Ó Dúill et al. 2006: 135) and is aware that this may limit the test task's validity. Nevertheless, virtually all best practice samples are designed for pairs of candidates. Likewise, he mentions the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF)* and even outlines some critical points. With regards to the KMK-Zertifikat, however, all there is to link it to the CEF is the test developers' claim. These two examples are symptomatic as well as indicative for the entire BLK study. The parts by Ó Dúill provide a good summary of the current state-of-the-art in language testing but appear to be completely unconnected and even irrelevant to the other parts which have practical implications.

2.2.3 Description and evaluation of the KMK-Zertifikat's item-types

Both of the items aiming to test the passive skills, i.e. listening and reading comprehension, respectively, contain multiple choice questions of a dichotomous type. However, even commonsensical considerations make clear that

> True/False or Yes/No items are generally unsatisfactory, as there is a 50% possibility of getting any item right by chance alone. In order to learn anything about a student's ability, it is necessary to have a large number of such items in order to discount the effects of chance. (Alderson, Clapham, Wall 2005: 51)

Obviously, 10 questions are not enough to rule out the possibility of students' answering correctly by simply guessing. This alone would be a problem big enough to label this test item as inappropriate for measuring a testee's listening/reading comprehension. However, there is another objection to these test items: in various cases, it is possible to arrive at the correct answer without having to read or listen to the text or by guessing. Simple commonsense and background knowledge suffice (cf. Alderson, Clapham, Wall 2005: 50 & 70 and Weir 1997: 41).

Reading Comprehension

Let us have a look at the 2005 reading comprehension test item for people being trained as office clerks (Bürokaufleute) at level B2. Our world and general knowledge is enough to answer several questions correctly without even having to look at the text. It is a known fact that India is not a leader in biotechnology and alternative energy. Furthermore, it is highly unlikely that specialists in India, which is in many respects still a developing country, would earn as much as in Western countries. The third question is even easier to answer. Living in a 'Western industrial country', who would fail to realise that managers are criticised by their own staff, who fear losing their jobs? Going on to the fourth

question, anyone who has had only the slightest experience with Indian culture will know that Indians are less likely to adhere strictly to an exact schedule or deadline than Germans. Judging from one's own experience helps to answer question number seven correctly and question number eight is simply redundant: who, regardless of their individual cultural background would ever claim that they were able to exclude the possibility of problems in the course of the business relations? Thinking of the high value less Westernised societies attach to the family, it is easy to get question nine right; and, though it seems almost too obvious and easy, there is a high possibility that statement number ten: "The Indian manager combines tradition and a dynamic work-ethos" is correct.

Only regarding question five and six, does background knowledge not help us. In one case, this is due to the fact that we cannot know Ms Jhaveri's opinion; in the other, the expected answer is incorrect, since there is no hint of it in the text. This still leaves us with at least an 80 % chance of guessing the correct answer without even having read the text!

Looking at the text, we wonder how it might possibly fit in with the claim of authenticity. For this item to be considered authentic in a test for vocational English in the office work place, we would expect it to be taken from the work environment. However, this is not the case. Rather, it resembles a newspaper article and only the vocabulary has to do with the specified topic.

In conclusion, we can state that this test item is designed unsatisfactorily in each and every respect (cf. 1.6.1) and therefore cannot be considered an appropriate tool for measuring a student's reading comprehension at all. When deciding to use a MCQ format, it is advisable to avoid dichotomous questions and present proper distractors.

Listening Comprehension

The Listening Comprehension consists of two parts. In the first part, the candidates listen to a conversation twice and have to answer ten Yes/No-questions. In the second part, they listen twice two a message on an answering machine and have to note down the name of the caller, his phone number and the order number. Both parts are again taken from the 2005 exam for people being trained as office clerks, this time at level B1.

Part 1:

When listening to the speech sample, one is immediately struck by the inauthenticity of the conversation presented and 1min 40sec into the speech sample, the listening comprehension item reaches its absolute low, in my opinion. At this point, it is evident that what we are witnessing is not a natural conversation, but rather a contrived text being read out aloud. Real conversation is normally characterised by being made up not

> of well-formed grammatical sentences, but rather short, clause like idea units: each about seven words long, of about two seconds duration, consisting of a single coherent intonation contour; these are usually strung together, or joined by conjunctions. Spoken language tends to have more non-standard features, such as dialect, slang, and up to date colloquialisms; and it also contains many disfluencies, such as fillers and hesitations [...], false starts, self corrections, afterthoughts, and so forth (Buck 1997: 66).

Now let us have a look at the sentence: "It is my job to redistribute the mail and prepare letters for my boss to sign." The first thing to notice is the relatively formal construction one would not expect in a normal conversation. Apart from this, putting the stress on "my" implies a contrast, as in: "My colleagues do not have any contact with customers. They write the PC programs we sell, whereas it is my job to deal with customer queries." However, in the speech sample, there is

no such contrast. Therefore, emphasising "my" sounds odd. Probably, the speaker has simply got tired of reading out a pre-constructed text. As a result, her intonation becomes monotonous. The sentence fragment "letters for my boss to sign" receives no stress at all, with the intonation continually falling[30]. The prosody, however, is not the only thing which is unnatural. Almost everything that would mark a normal everyday conversation is lacking. There are virtually no emotives, colloquialisms, etc. Instead, relatively formal expressions, such as "moreover", feature prominently in the speech sample. Interestingly, the male speaker includes a certain amount of hesitation and fillers into his speech, whereas his female partner does not.

What adds to the perception of the sample's artificiality is its crooked logic. Thus, right at the beginning, when the male speaker informs us that, sometimes, he has to work longer, the female speaker replies by saying: "Sounds good!", adding a somewhat ironic undertone[31] which is unlikely to be intended. In addition to this, many aspects seem to have been included only for the sake of using apparently work-related vocabulary. While it is true that conversations are unlikely to develop linearly, I find the way in which the topic of the conversation is changed highly implausible. Asked about the software they use, one of the speakers answers briefly by saying: "We use a range of Microsoft Office 2000 programs" (02:38-02:42). Immediately afterwards, he starts a lengthy monologue about how his company books their business trips. Since both speakers are supposed to be Personal Assistants to their bosses, we would expect that both of them are

[30] Another example of the strange intonation patterns can be found in 00:47 – 00:51. Throughout the entire phrase: "my boss has also three other people on his team", the pitch is unexpectedly high and even rises at the end.

[31] I am well aware that this paradox is explained later on, as the female speaker has to work much longer than the male. However, I still consider the formulation unlikely to occur in any natural conversation. Rather than saying "sounds good", we would expect a mediating expression such as "Well, compared to what I have to do, that sounds really good!" or "Not too bad, compared to my working situation!"

familiar with the individual tasks included in 'making travel arrangements'. Nevertheless, the male speaker spells it out for his female counterpart. Now she knows for sure that it "also includes booking flights, hotels, rental cars and entering all details into electronic calendars" (02:59 - 03:06). The piece of information he gives just before this statement is probably one of the most indicative: "We book travel online, with our electronic travel center, and view travel plans virtually" (02:50 - 02:55). Like the listing mentioned above, this is of little informative value to the other speaker and leaves us wondering why it is added to the conversation. It seems that the authors simply wanted to test a particular set of vocabulary and so constructed a conversation containing all these words. This claim is supported by many examples. Just think of the formal nominal construction in "since the merger with an American company" (03:14 – 03:17) or the female speaker's slight hesitation in 03:22 - 03:23, which somewhat stresses the following word "in-house management". It appears that the authors' supposed preoccupation with vocabulary issues has led to an inauthentic conversation in terms of content as well as prosody. Apart from that, the authenticity of the very task itself is questionable.

As already mentioned in the introductory remarks, another problem with this listening comprehension item is the dichotomous question format. This item fulfils neither the requirements for authenticity nor can it be regarded as a valid measure of the testees' listening skills. The first part of this listening comprehension must therefore be dismissed as completely unsatisfactory.

Part 2:

In the speech sample for the second part of the listening comprehension, none of the problems inherent in the first part occur. The message is delivered at normal speed; the content and task are authentic. The candidates are to fill in the information on a form which is provided on the answer sheet.

All in all, the second part is an authentic and appropriate task for measuring listening skills.

2.2.3 Description and evaluation of the KMK-Zertifikat's item types

Writing

The writing subtest consists of two parts: the first part is made up of two items which are to test the candidates' productive skills, the second part comprises one item testing their mediating abilities. The items evaluated are taken from the 2006 exams for office clerk trainees at level B1 for the first two items and B2 for the last item.

Part 1:

The first tasks demands that the testees write a letter. The content of the letter as well as its structure is given in German. Therefore, it is to be doubted that what this item assesses is someone's writing skills. Clearly, this task does not assess whether a candidate is able to "write extended stretches of meaningful, literate discourse in the language being evaluated" (Cumming 1997: 52) because this includes skills such as structuring the text and taking into account appropriate content as well as appropriate style. Since structure and content are already specified and the problem of using the appropriate style can, to a certain extent, be evaded by translating the German text into English, what is being tested is arguably translation - with only minimal writing skills being necessary. For measuring writing, this test item is therefore inappropriate although it may be useful for assessing translating abilities.

The second task is basically a shorter variant of the first item. This time, the examinees are to compose an e-mail instead of a letter. Otherwise, the tasks are very similar. Again, all information, structural as well as content-wise is given. Just like the first item, the second one is therefore inappropriate for determining writing skills. However, it could be justified as a means of assessing translation skills.

Part 2:

The third task, 'mediation', asks the student to paraphrase the German text in English. It does not have to be a literal translation. In fact, the test developers maintain that "generally, translation or interpretation should not be the focal points" (Ó Dúill et al. 2005: 73, my translation). However, the main difference they mention for distinguishing between 'mediation' and 'translation' is the situational embedding. Whereas a translation tasks is seldom accompanied by additional information outlining the situation, this is compulsory for a mediation task, they say (cf. ibid.). However, it is questionable whether this really has an influence on the testees. Furthermore, the only indication that the candidate does not have to provide a literal translation is the subtle difference between the words "übertragen" and "übersetzen". It could well be that testees miss this hint and consider the task a translation.

Although only the second part purports to be measuring translation or mediation skills, all three item types are appropriate for measuring these abilities. However, the items in part one are not adequate for assessing writing, which they are supposed to do. In terms of authenticity, all tasks score high. Each of them represents jobs they could be expected to carry out when working as an office clerk later on in their career.

As far as scoring is concerned, a common rating scale is available, which could help in ensuring common standards and high inter-rater reliability. However, it has to be noted that a common rating scale alone, even if it is based on the CEF does not generate reliability and validity by itself. Unfortunately, though, this is what the developers of the KMK-Zertifikat seem to think. Engelhart's statement that the "descriptors [...] are based on the formulations of the Common European Framework of Reference for Languages: Learning Teaching and Assessment [...and] are **thus** reliable and valid"(Ó Dúill et al. 2005: 125, my translation. emphasis added) indicates a substantial lack of understanding as to what reliability and validity mean and how they are established and proved.

2.2.3 Description and evaluation of the KMK-Zertifikat's item types

Speaking

As the oral examinations for the KMK-Zertifikat in Baden-Württemberg are planned by the individual schools, no samples from former examinations were available to me. The sample task evaluated is taken from the best practice examples in Ó Dúill et al. (Ó Dúill et al. 2006: 179) designed for office clerk trainees at level B1.

The testing takes place in pairs and is arranged as a role play. Both candidates are assigned the role of an employee who has to plan the furnishing of the company's customer information bureau together with an English-speaking colleague. Each of the examinees is to assume that their partner is the English speaking colleague. On their role play cards, each of the candidates is provided with information about the office as well as with her character's preferences regarding the furnishing. In the ensuing conversation, the candidates are to show that they can communicate effectively by exchanging the information specified on their respective role-play cards and by negotiating their differing views regarding the furnishing.

Apart from arranging the exam as a group assessment, the task is well designed. The roles and situation can reasonably be considered authentic.

Regarding the scoring process, however, the situation is similar to the tasks writing. Although there is a common rating scale, no other measures have been taken to ensure inter-rater reliability.

2.3 Conclusion: summary – perspectives

In this conclusion, I will sum up the findings of the above sections and compare the two tests in these respects. I will also hint at the current developments concerning both tests.

2.3.1 Summary

In the preceding sections it has become clear that for both tests, there are good as well as bad aspects. As for validity and reliability, it has been shown that the TOEIC test is highly reliable and valid in terms of listening and reading comprehension. The KMK-Zertifikat, on the other hand, has not even had common rating standards until recently and no action whatsoever has been taken to estimate, let alone ensure, reliability and validity. Moreover, the reading comprehension and the greater part of the listening comprehension section of the KMK-Zertifikat are absolutely inappropriate. The speaking section is better, although there are also concerns regarding the group format which might bias in favour of more proactive students and against shyer candidates or candidates whose performance is restricted and hindered by their partners.

Both tests include items which assess skills other than those they are purported to test. Regarding the TOEIC test, this is the case with the first two parts of the Reading Comprehension section which assess grammar rather than reading comprehension. As for the KMK-Zertifikat, this is the case with the items supposedly testing writing, when in fact assessing translation. However, when thinking of the tests as tests of general English ability, this may be excusable.

This brings us back to the statements and claims which are made in the respective marketing material for the tests. Regarding both tests, there are some claims which lack substance. As for the TOEIC, this is particularly true for the claim that, supposedly, "the TOEIC test does provide an indirect measure of speaking and writing" (ETS. *TOEIC User Guide*: 8). According to the test developers, studies "have confirmed a strong link between TOEIC results and oral proficiency.

Smaller studies have shown a similar link with writing skills" (ibid.). While it is problematic to say that these claims are wrong – if only because there is no definition informing us how strong a link needs to be to be called a 'strong link' – I would maintain that they are misleading. Test users such as companies or academic institutions may not be aware of the amount of uncertainty involved in these putatively strong links and may use test results for unwarranted purposes. Particularly in high-stakes decisions, this is extremely worrying. Since there has not been any study to evaluate the TOEIC test's predictive validity yet, one should also be wary of statements asserting that "TOEIC scores can indicate whether an employee will be able to work and interact successfully if posted to an English-speaking country" (ETS. *TOEIC User Guide*: 7). Likewise, using the TOEIC for "monitoring individual or group progress" or for evaluating "the effectiveness of [... a] program in improving students' English language proficiency" (ibid.) can be problematic and lead to negative washback. Using the TOEIC for these purposes may give rise to unrealistic expectations on the students' part. Despite the fact that a minimum of "at least 100 hours of language training is usually required before students are able to demonstrate a real increase in TOEIC scores" (ETS. *TOEIC User Guide*: 11), students enrolled in language courses – which only rarely contain 100 hours of language instruction – may expect TOEIC score gains in significantly less time. As Cunningham has shown, this is possible, although score gains thus obtained may have little to do with improved communicative competence and rather reflect improved test-taking skills (cf. Cunningham 2002).

As for the KMK-Zertifikat, there are also some claims which need to be handled with caution. Above all, this is true for the claim that the KMK-Zertifikat "is recognised in 41 member states of the Council of Europe" (LIS 2006. my translation). All this statement is based on is the putative link to the CEF, which, according to the test's developers, warrants Europe-wide recognition. However, even in Germany, the KMK-Zertifikat's popularity with employers is questionable. Ó Dúill quotes an unpublished study conducted with 20 companies

in Lower Bavaria, according to which 24% of those interviewed would preferentially hire a candidate with KMK-Zertifikat. In my opinion, this is quite a low figure and does not speak in favour of the KMK-Zertifikat's acceptance as a valuable qualification. This view is reinforced by the findings of my own study, which reveal that only 7 % of teachers[32] know the KMK-Zertifikat exists at all (cf. chapter 3.3). The KMK-Zertifikat's link to the CEF has also been shown to be a claim which lacks evidence. For the TOEIC test, on the other hand, this link has been verified empirically, although one must be careful not to over-emphasise such links. Regarding the degree of popularity with teachers, however, the study shows that the TOEIC is known even less than the KMK-Zertifikat. Only 2.8% had heard of the TOEIC test before (cf. chapter 3.3). On the other hand, it is a fact that the TOEIC is used by many institutions and businesses world-wide, whereas the need for the KMK-Zertifikat is less clear. Its being intended as "rein schulisches Zertifikat" (Janssen: 2001: 4) which is designed according to the school curricula, as well as its being administered and awarded by the schools may raise the question in which way it adds to the actual school-leaving certificate. In fact, the wish to increase learners' motivation, i.e. the wish to create positive washback seems to have been the main reason for implementing the certificate.

All in all, the TOEIC can be considered a useful tool for assessing someone's listening and reading comprehension. It may also be helpful for estimating grammatical competence. It is strongly advised that the TOEIC test be not used for any other purpose, since this may result in incorrect conclusions.

The KMK-Zertifikat on the other hand may be appropriate for measuring speaking and translating skills, although no data substantiating claims of validity and reliability is available. As a measure for listening and reading comprehension as well as writing skills the KMK-Zertifikat is inappropriate.

---

[32] It has to be said, though, that the respondents teach at grammar schools (Gymnasien), whereas the KMK-Zertifikat aims at pupils attending vocational schools (Berufsschulen).

2.3.2 Perspectives

During the time this paper was being written, important changes have taken place. Some of the critical points mentioned in the preceding sections have now been addressed. It is now possible to take optional speaking and writing tests to accompany the TOEIC listening and reading comprehension. Furthermore, the section on 'error recognition' has been substituted by a 'text completion' exercise which is similar to the part on 'sentence completion', but differs in that it provides more context material.

Regarding the KMK-Zertifikat, applying the criteria Charles Alderson outlined in a presentation (Alderson 2006, see appendix) held to the test's developers leads to the conclusion that the current practice of administering the KMK-Zertifikat is unprofessional. However, the test developers have been made aware that

> a lot has yet to be done, if the KMK-Fremdsprachenzertifikat wants to offer a product which is comparable to those of the big testing agencies [...]. As for the KMK-Fremdsprachenzertifikat, so far, procedures such as the pre-testing of items or the creation of item banks have not been realised. So far, the statistical evaluation of the [...] exams has been underdeveloped (Ó Dúill et al. 2006: 207. my translation).

Therefore, there is hope that the KMK-Zertifikat can improve in the future. Yet, whether it does, in fact, improve is not only a matter of knowledge concerning test and item design, but also a political question. Since the test is developed by teachers and administered by schools, it is basically state funded. Now, it is up to politicians to decide how much value they want to attribute to language learning and testing. It is up to them to decide how much money they are prepared to spend on improved educational policies. For "[g]ood tests and assessment, following European standards, cost money and time. But [... b]ad tests and assessment, ignoring European standards, waste money, time and LIVES [sic!]" (Alderson 2006: 43).

3 Survey

In the following evaluation of a survey concerning language tests among teachers at German grammar schools (Gymnasien), I will investigate whether there exist correlations between outside factors such as the location of the school, number of students or the inclusion of a statement of commitment to foreign languages within the mission statement and the interest of teachers in language tests, viz. their inclination to promote or even administer such tests.

3.1 Method

The study was planned and conducted as part of an internship with Sprachenmarkt.de, a subsidiary to LearnBiz.com. As a provider of language travels and other language related services such as language tests, they had naturally developed an interest in the issue of external assessment at schools.

First, a questionnaire (see appendix) aiming at the greatest possible neutrality and least possible bias was designed. To ensure the proper functioning of the questionnaire, it was tried out and tested by sending it out to non-target groups[33]. Thus, it could constantly be improved until the final version was used to actually conduct the study in September and October 2006. In a second step, all grammar schools in Baden-Württemberg were contacted by telephone to enquire their readiness to take part in the study. In some cases, however, it was impossible to reach the people in charge, in others, they preferred not to take part in the study. Finally, the questionnaires were sent out to those 230 schools which agreed to take part, either by mail, email or fax, according to choice. Of the 230 questionnaires which were sent out, 142 were returned and analysed.

[33] Non-target here refers only to the location of the respondents rather than to other criteria.

3.2 Hypotheses

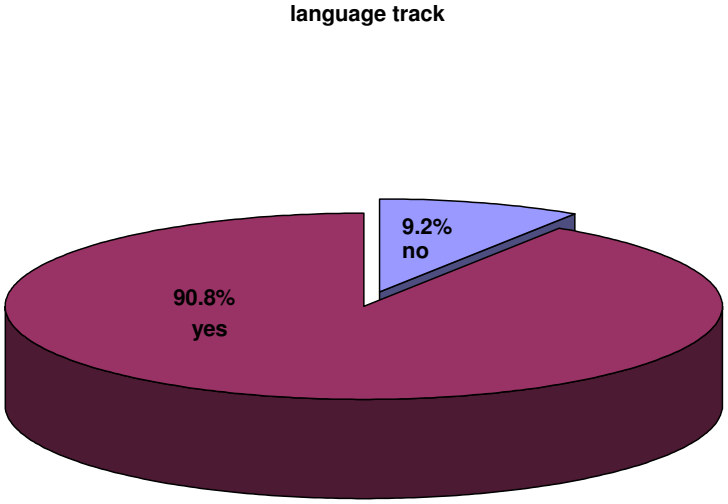
Before analysing the data it is worth setting forth the hypotheses with which the topic was approached. As said already, the knowledge of foreign languages, especially English, plays an important role in today's society – not least in view of economic necessities. Therefore, a high amount of schools offering a language track with the opportunity to learn more than the compulsory two foreign languages (which is known as 'Sprachenzug' [language track] as opposed to the more science oriented 'naturwissenschaftlicher Zug' [science track]) could be expected. Likewise, the number of schools prioritising languages in their mission statement (Schulprofil/Leitbild) could be assumed to be similarly high. Furthermore, I supposed that both of the aforementioned criteria positively affected the willingness to offer language tests. Out of personal experience, however, I did not believe that there would be too high a percentage of schools actually implementing external assessment such as language tests already.

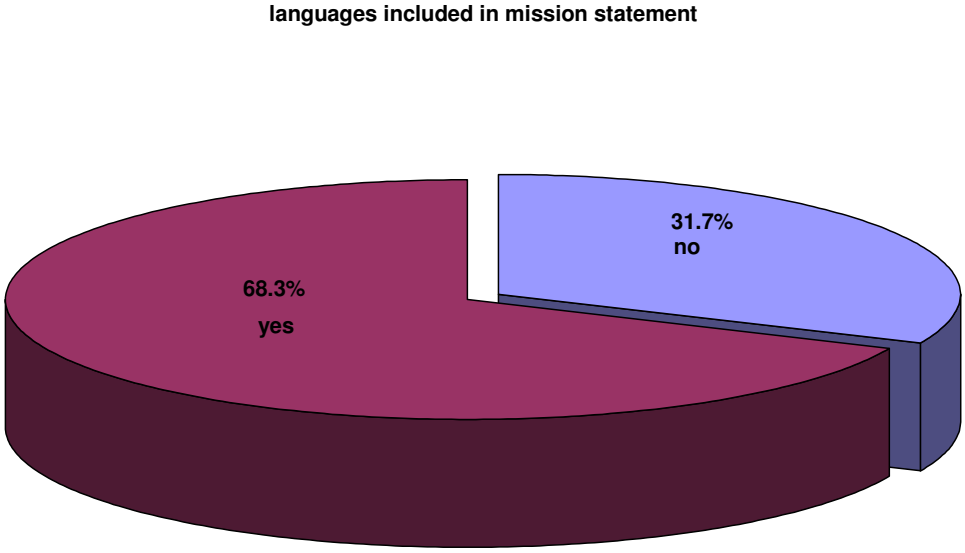
It was presumed that a high number of pupils triggered an increased readiness to administer extra-curricular language tests. Moreover, due to the heightened awareness of problems on the job market in urban surroundings, the motivation to offer additional possibilities for certification was supposed to deteriorate in more rural areas. Apart from that, this motivation was also expected to correlate with teachers' level of familiarity with the various different language tests and certificates. As for the decision for or against a particular test, the following criteria were expected to be important:

- Price
- Time / effort needed for preparation and administration
- International recognition
- Recognition as fulfilling universities' admission criteria
- Scientifically sound basis
- Linkage to the Common European Framework
- Popularity

3.3 Analysis

As can be seen in the charts and diagrams, the hypotheses concerning the mission statement and the language track can easily be verified:

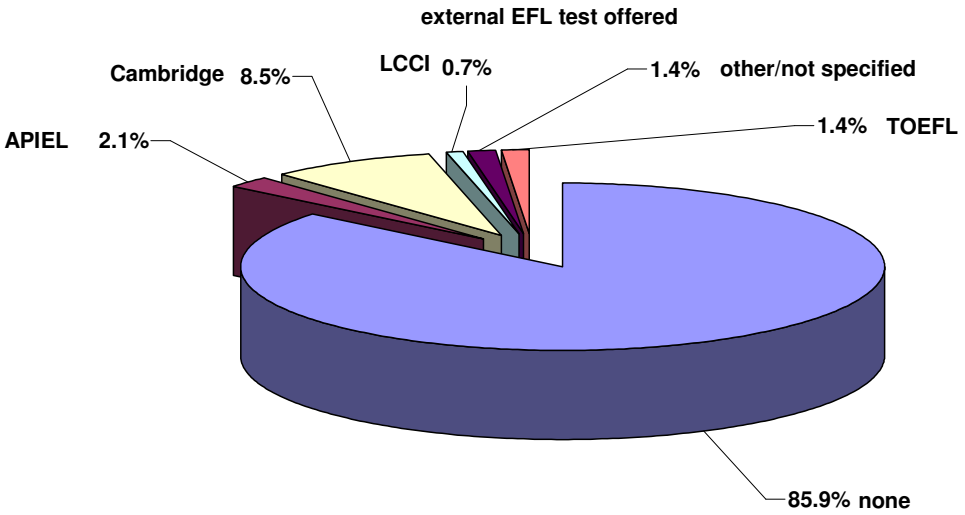




In fact, an overwhelming majority of 90.8 % of schools offer pupils the opportunity to learn more than two foreign languages and 68.3 % even include a commitment to foreign languages in their mission statement. Since only 7 % of the respondents were totally opposed to the idea of offering external EFL tests at their school, the data does not suffice to support or refute the hypothesis that having a language track or including languages in the mission statement influences the willingness to offer language tests. Likewise, no positive link between the availability of a language track and the availability of external

language assessment could be established. However, there seems to be a link between the inclusion of languages in the mission statement and the availability of language tests. The likelihood that language tests are offered at schools including a commitment to languages in their mission statement is almost twice as high as at other schools (16.5 % vs. 8.9 %). Not considering the phased out APIEL test in the calculation yields even clearer results, namely 15.5 % vs. 4.4 %.

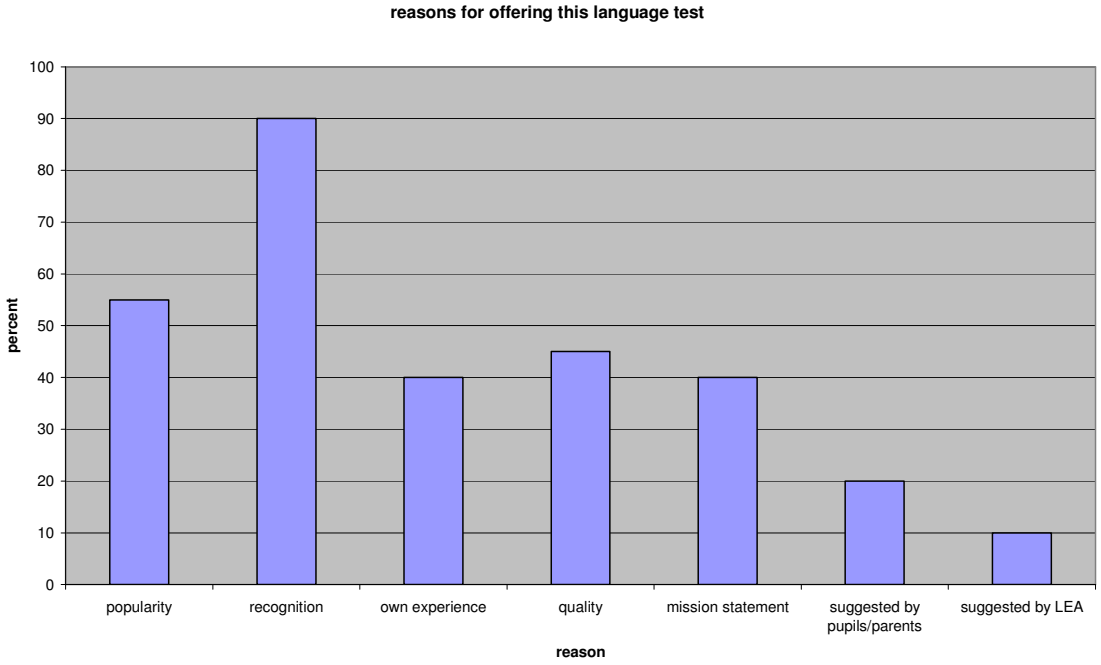
Those schools which provide the possibility of obtaining an additional certificate make up only 14.1 percent. Regarding the actual certificates which are offered, Cambridge (including all the tests from PET over BEC and CAE to IELTS) with 8.5 % takes the lead. APIEL[34] follows with 2.1 %, while 1.4 % opt for the TOEFL. Only 0.7 % choose the LCCI exams. Though designed as a simple question, there were multiple answers in two cases. Both of them would have to be included in the 'other/not specified slice, as one featured the SAT and the other a national language competition rather than a language test. These findings are summarised in the following pie chart:



[34] This is a rather odd result, as, to my knowledge, APIEL has been phased out in Germany since 2003.

3.3 Analysis

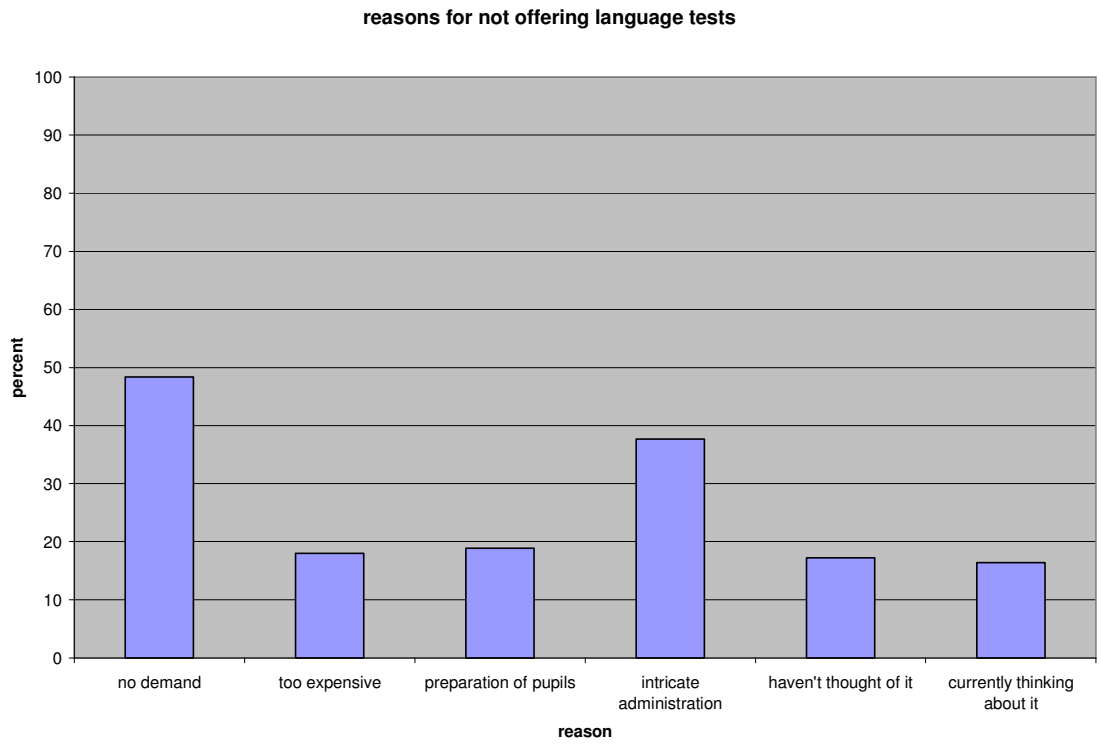
The reasons for having chosen a particular test over another are displayed in the following diagram:



While 55 % mentioned 'popularity' as a reason for introducing this particular test, an overwhelming, yet not surprising majority of 90 % claimed that the test's 'recognition' was an important factor in the decision. Although the results for the teachers' own experience and the commitment to languages through the mission statement could be expected to range somewhere in the middle – and in fact, turned out to figure in the low mid-range, i.e. 40 %, respectively – and one would not have imagined suggestions by pupils/parents or the Local Education Authority (LEA) to rank higher than they did (20 % and 10 %, respectively), only 45 % referred to the quality of a test when giving reasons for their choice. This, to me, seems like a rather disturbing fact, as it has us believe that the substantial majority of teachers attributes only little value to quality as opposed to, say, recognition. Bearing in mind that a test's recognition does not necessarily correlate with its quality, emphasising and pointing out the importance of a test's quality must be a key concept in all future activities related to language assessment at school.

Of those schools currently not offering any additional language assessment, only 10 could not imagine administering a supplementary language test under any circumstances. All others were open to the idea of offering extra-curricular

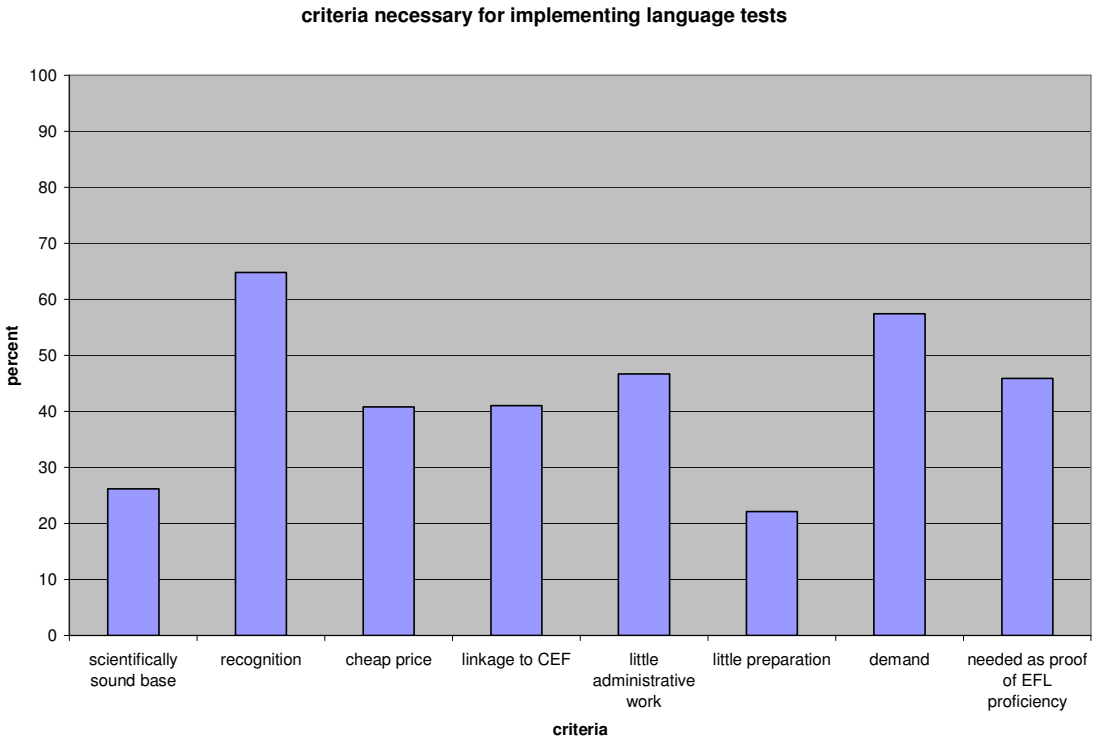
language testing, provided certain criteria are met. However, this is put into perspective again when considering that 48.4 % of teachers do not see any demand for an extra language test. The evaluation of the reasons for not offering language certificates at school also shows that there is still quite a substantial number of schools (namely 17.2 %) where teachers have never thought about implementing additional language assessment. On the other hand, 16.4 % claim to be already thinking about introducing an extra certificate. For surprisingly few teachers, the price played an important role. Only 18 % mentioned it as a reason against introducing a language test. Teachers were slightly more worried about the additional work load they might have to expect. 18.9 % voiced concerns regarding the complex and time-consuming preparation of the pupils for the test. Some (37.7 %) also objected to the intricate administration process. All of these findings are summed up in the following diagram:



Regarding the criteria which would have to be met in order to allow for the administration of language tests at school, again, the potential certificate's recognition, with 64.8 %, ranks highest, just above the demand as voiced by pupils and parents with 57.4 % and the possible use as a proof of the pupils' EFL

proficiency. With 41 % teachers attribute even slightly more value to the linkage to the CEF than to a cheap price. Whereas it is more than understandable that teachers are unlikely to promote a test which would make much more work on their part necessary, it seems laudable that less than half (46,7 %) of the teachers mention administrative efforts as influencing their decision. As for the preparation of the pupils, this figure is even lower, namely 22.1 %.

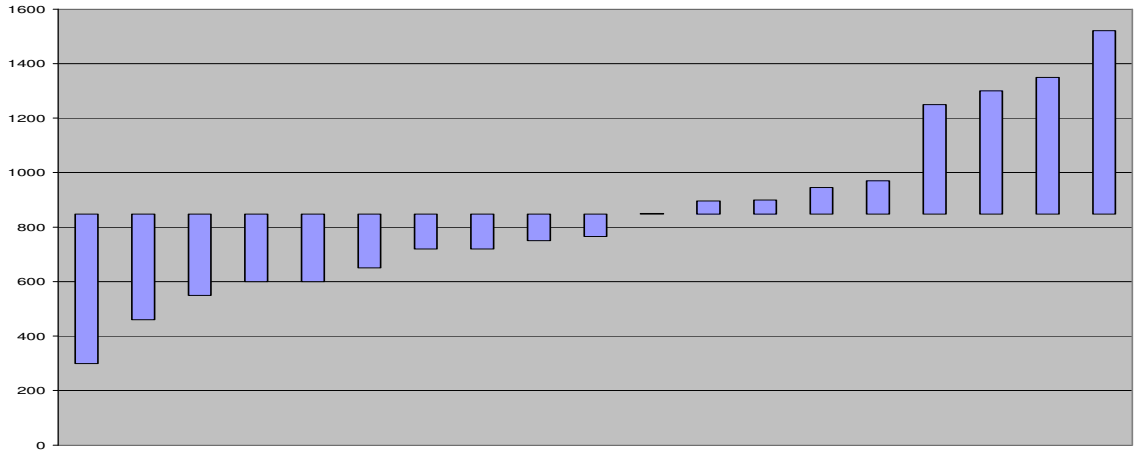
However, it is rather distressing that qualitative criteria such as a scientifically sound base are among the least important factors. This is even more disturbing if we consider the emphasis which is laid on evasive and sometimes subjective notions as 'recognition' or 'popularity'.



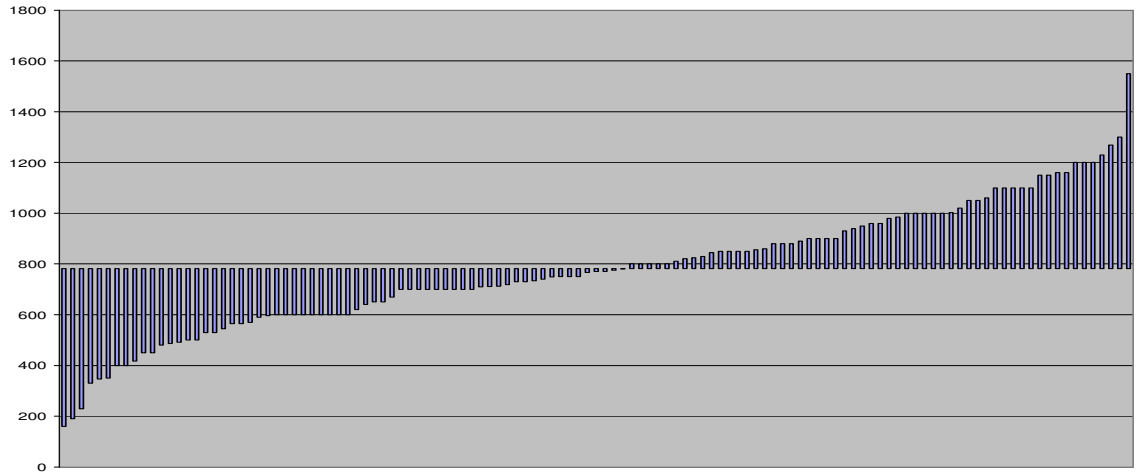
As for the hypotheses concerning the correlation between the number of pupils at a school and the likeliness to offer a certificate, they seem to have been wrong. A correlation between the size of a school and the probability that this school offers a certificate could not be established, which is revealed in the following

diagrams. The first chart displays the size of schools administering a language test, the second the size of those without. Whereas the size of schools with a certificates ranges from 300 to 1521 pupils with a mean of ~847, the size of schools which do not offer certificates ranges from 160 to 1550 pupils with a mean of ~782. The average size of all schools taking part in the survey was ~791[35].

size of schools offering EFL tests 1



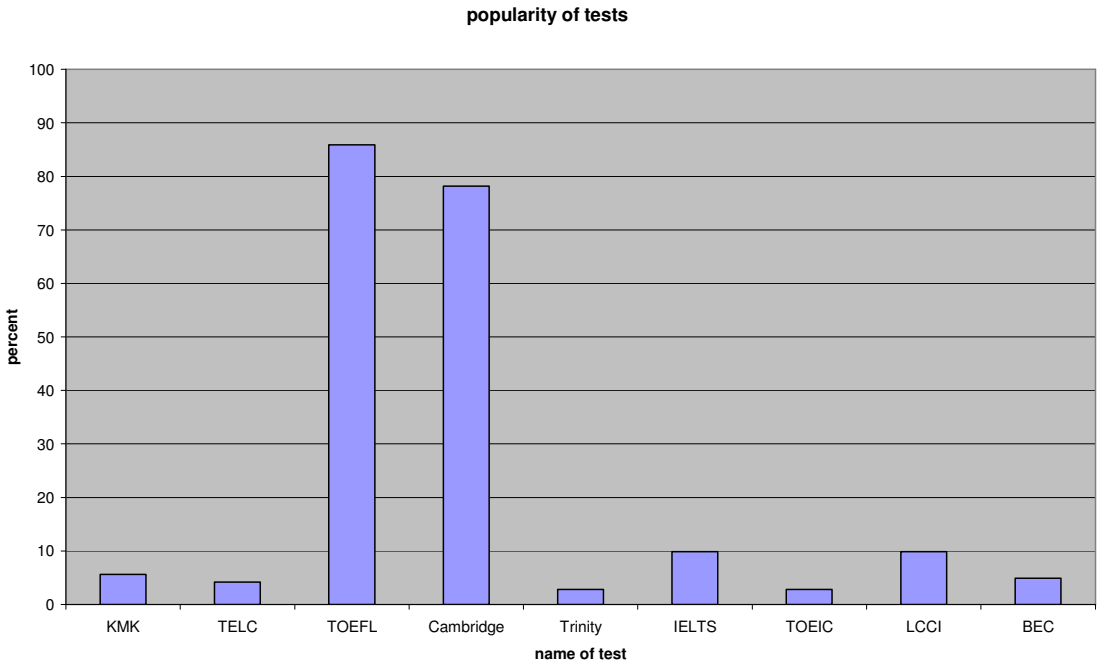
size of schools not offering EFL tests



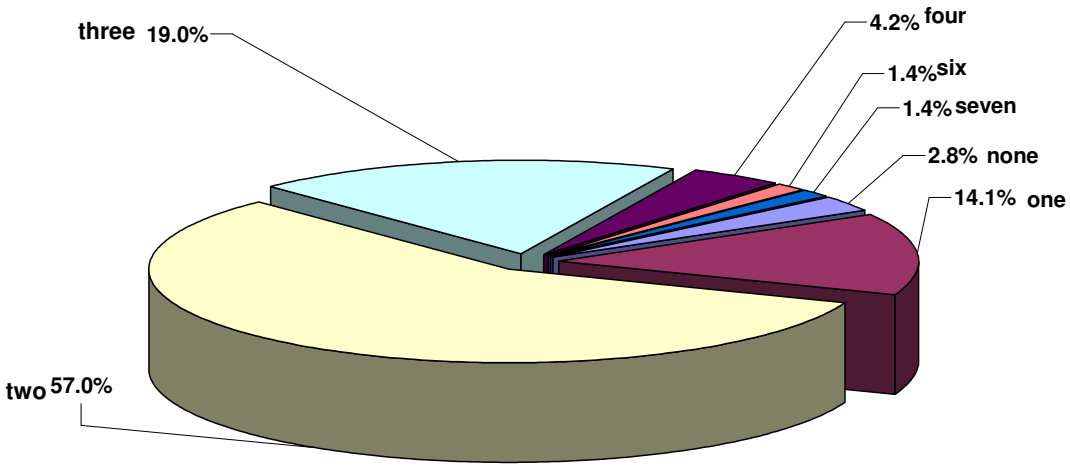
[35] Of the 142 schools returning the questionnaire, 2 (one administering a language test) did not answer this question and therefore had to be neglected in this particular statistic.

Likewise, there was no discernable link between the availability of a language test and the location of the school. Both, rural and urban schools offer about the same possibilities regarding language testing.

Regarding the popularity and familiarity of teachers with various tests, the TOEFL test takes the lead with 85.9 %, closely followed by the Cambridge Certificates which were known by 78.2 %. The situation for the other tests is less favourable. 9.9 % were familiar with the LCCI exams and the IELTS test, 5.6 % with the KMK-Zertifikat and 4.2 % with the TELC examinations. 4.9 % of the respondents knew the BEC and with a meagre 2.8 %, the TOEIC and Trinity tests take the bottom end of the scale.



Of the certificates mentioned, the majority of the respondents (57 %) knew two. 19 % were familiar with three, 14.1% with one and 4.2% with four. 1.4 % recognised six and seven tests, respectively. Only 2.8 % did not know any of the language tests mentioned. These figures are reflected in the following pie chart:



The assumption that the availability of language tests was influenced by the amount of language tests the teachers knew could not be substantiated.

Of the above findings, the fact that the respondents placed remarkably little value on the tests' quality is particularly distressing. As has become clear, they relied mainly on a test's recognition. Therefore, educating teachers about the importance of quality and the criteria such as reliability and validity which are essential elements necessary to determine a test's quality must be a core issue in all actions relating to language testing at schools. The need for such further training could also be seen in chapter 2. Again, however, whether this training will be made possible, whether this education will take place is due to political decisions. As all training costs money, politicians have to decide if they really want to improve educational policies and how much worth this improvement is to them.

Bibliography

- Alderson, Charles, Clapham, Caroline & Wall, Diane. 2005. *Language Test Construction and Evaluation*. Cambridge: CUP. 9th ed.

- Alderson, Charles & Wall, Diane. 1993. 'Does washback exist?'. *Applied Linguistic* (14). 115-129.

- Bachman, Lyle. 1990. *Fundamental Considerations in Language Testing*. Oxford: OUP.

- Bachman, Lyle & Palmer, Adrian. 1996. *Language Testing in Practice*. Oxford: OUP.

- Bachman, Lyle. 1997. 'Generalizability Theory'. In: Clapham, C. & Corson, D. (Eds.). 255-262.

- Bachman, Lyle & Eigner, Daniel 1997. 'Recent Advances in Quantitative Test Analysis'. In: Clapham, C. & Corson, D. (Eds.). 227-242.

- Biesemann et al. 2005. *Fremdsprachenzertifikate in der Schule*. Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen.

- Baker, David. 1989. *Language Testing – A Critical Survey and Practical Guide*. London: Edward Arnold.

- Banerjee, Jayanti & Luoma, Sari 1997. 'Qualitative Approaches to Test Validation'. In: Clapham, C. & Corson, D. (Eds.). 275-287.

- Brown, James Dean & Hudson, Tom. 2002. *Criterion Referenced Language Testing*. Cambridge: CUP.

- Buck, Gary. 1997. 'The Testing of Listening in a Second Language'. In: Clapham, C. & Corson, D. (Eds.). 65-74.

- Chapelle, C.A. 1999. 'Language Testing: Methods'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 721-724.

- Clapham, C. & Corson, D. (Eds.). *Encyclopedia of Language and Education. Volume 7: Language Testing and Assessment*. 1997. Dordrecht: Kluwer Academic Publishers.

- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Cambridge: CUP. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf. (accessed 08/06/2007).

- Council of Europe. 2003. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF)* – Manual (Preliminary Pilot Version). Language Policy Division, Strasbourg. http://www.coe.int/T/DG4/Portfolio/documents/Manual%20for%20relating%20Language%20Examinations%20ot%20the%20CEF.pdf (accessed 08/06/2007).

- Cumming, Alister. 1997. 'The Testing of Writing in a Second Language'. In: Clapham, C. & Corson, D. (Eds.). 51-63.

- Cunningham, Cynthia R. 2002. *The TOEIC Test and Communicative Competence: Do Test Score Gains Correlate With Increased Competence? A preliminary study*. University of Birmingham: MA thesis [unpublished]. http://www.cels.bham.ac.uk/resources/essays/Cunndiss.pdf (accessed 06/07/2007).

- Davidson, F.; Turner, C.E. & Huhta, A.: 'Language Testing Standards'. In: Clapham, C. & Corson, D. (Eds.). 303-311.

- Davies, Alan; Hamp-Lyons, Liz & Kemp, Charlotte. 2003. 'Whose norms? International proficiency tests in English'. *World Englishes* (22/4). 571-584.

- ETS. 2000. *TOEIC Form ST-00* (Sample Test). Princeton: ETS.

- ETS. 2006. *TOEIC Form ST-05* (Sample Test). Princeton: ETS.

- Ferguson, Charles A. 'Foreword'. In: Kachru, Braj B. (Ed.). *The Other Tongue – English across Cultures*. Urbana/Chicago/London: University of Illinois Press.

- Fulcher, Glenn & Davidson, Fred. 2007a: *Language Testing and Assessment*. London: Routledge.

- Fulcher, Glenn. 1997. 'The Testing of Speaking in a Second Language'. In: Clapham, C. & Corson, D. (Eds.). 75-85.

- Fulcher, Glenn. 2004. 'Are Europe's tests being built on an unsafe framework?'. *The Guardian Weekly TEFL Supplement*, 18. March 2004.

- Fulcher, Glenn. 2007. 'Universities undermine their own foundations'. *The Guardian Weekly TEFL Supplement*, 13. April 2007.

- Fulcher, Glenn & Davidson, Fred. 2007b: 'The Common European Framework of Reference (CEFR) and the design of language tests: A Matter of Effect'. *Language Testing* (40). 231-241.

- Gilfert, Susan. 1996. 'A Review of TOEIC'. In: *The Internet TESL Journal*. http://iteslj.org/Articles/Gilfert-TOEIC.html (accessed 19/03/08).

- Hamp-Lyons, Liz. 1997. 'Ethics in Language Testing'. In: Clapham, C. & Corson, D. (Eds.). 323-333.

- Hirai, Michihiro. 2002. 'Correlations between active skill and passive skill test scores'. In: *Shiken: JALT Testing &Evalutaion SIG Newsletter*(6/3). http://jalt.org/test/hir_1.htm (accessed 19/03/08).

- Kelly, R. 1999. 'Foreign Language Testing'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 689-695.

- Kniffka, Gabriele. 2003: 'Prüfen und Bewerten'. In: Bausch, Karl-Richard; Christ, Herbert & Krumm, Hans-Jürgen (Eds.): *Handbuch Fremdsprachenunterricht*. 4th ed. Tübingen: A. Francke Verlag.

- Kunnan, A.J. 1999. 'Language Testing: Fundamentals'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 707-711.

- Lado, Robert. 1965. *Language Testing – The Construction and Use of Foreign Language Tests*. London: Lowe and Brydone. 4th ed.

- Landesinstitut für Schulentwicklung (LIS). 2006: *Das KMK-Fremdsprachenzertifikat*. http://www.ls-bw.de/beruf/pruefungen/kmk/zertifik.html (accessed 24/03/2008).

- Lynch, Brian K. & Davidson, Fred. 1997. 'Criterion Referenced Testing'. In: Clapham, C. & Corson, D. (Eds.). 263-273.

- McNamara, Tim. 2000. *Language Testing*. Oxford: OUP.

- McNamara, Tim. 1997. 'Performance Testing'. In: Clapham, C. & Corson, D. (Eds.). 131-139.

- McNamara, Tim. 1999. 'Language Testing: Users and Uses'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 724-728.

- Morfeld, Petra: 'Sprachenzertifikate'. In: Bausch, Karl-Richard; Christ, Herbert & Krumm, Hans-Jürgen (Eds.): *Handbuch Fremdsprachenunterricht*. 4th ed. 2003. Tübingen: A. Francke Verlag.

- Moritoshi, Paul. 2001: *The Test of English for International Communication (TOEIC): necessity, proficiency levels, test score utilisation and accuracy.* The University of Birmingham.
  http://www.cels.bham.ac.uk/resources/essays/Moritoshi5.pdf
  (accessed 06/07/2007).

- North, Brian. 2004: 'Europe's framework promotes language discussion, not directives'. *The Guardian Weekly TEFL Supplement*, 15. April 2004.

- Norton, Bonny: 'Accountability in Language Assessment'. In: Clapham, C. & Corson, D. (Eds.). 313-322.

- Ó Dúill et al. 2005/2006: *[eu] KonZert - Entwicklung und Umsetzung eines Evaluationskonzeptes für die KMK-Fremdsprachenzertifikatsprüfungen zur Sicherung der Vergleichbarkeit der Standards*. ISB, Staatsinstitut für Schulqualität und Bildungsforschung, Bayern; Behörde für Bildung und Sport, Hamburg; Thüringer Kultusministerium, Erfurt; Fachhochschule Rosenheim, Fachbereich Allgemeinwissenschaften.
  http://www.beruflicheschulen-modellversuche.de/showmv.php?i=3
  (accessed 14/04/2008).

- Oller, John W. Jr. & Perkins, Kyle. 1978: *Language in Education – Testing the Tests*. Rowley, Massachusetts: Newbury House Publishers Inc.

- Popper, Karl R. 1989. *Logik der Forschung*. Tübingen: Mohr. 9th ed.

- Quine, W. V. 1970. 'On Popper's Negative Methodology'. In: Schilpp, P. A. (ed.). *The Philosophy of Karl Popper*. 1974. La Salle, Illinois: The Open Court Publishing Company. 218-220.

- Rea-Dickens, Pauline. 1997. 'The Testing of Grammar in a Second Language'. In: Clapham, C. & Corson, D. (Eds.). 87-97.

- Schärer, Rolf. 2003. 'Sprachenportfolio'. In: Bausch, Karl-Richard; Christ, Herbert & Krumm, Hans-Jürgen (Eds.): *Handbuch Fremdsprachenunterricht*. Tübingen: A. Francke Verlag. 4th ed.

- Shohamy, Elana. 2001 *The Power of Tests – A Critical Perspective on the Uses of Language Tests*. Edinburgh: Pearson Education Ltd.

- Shohamy, Elana. 1997. 'Second Language Assessment'. In: Tucker, G.R. & Corson, D. (Eds.). 141-149.

- Shohamy, Elana. 1999. 'Language Testing: Impact'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 711-714.

- Smith, K. 1999. 'Language Testing: Alternative Methods'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 703-706.

- Spolsky, Bernard & Asher, R.E. (Eds.): *Concise Encyclopedia of Educational Linguistics*. 1999. Oxford: Elsevier Science Ltd.

- Spolsky, B. 1999. 'Language Testing'. In: Spolsky, Bernard & Asher, R.E. (Eds.). 695-703.

- Spurling, Steven. 1987. 'The Fair Use of an English Language Admissions Test'. *The Modern Language Journal* (71). 410-421.

- Tannenbaum, Richard J. & Wylie, E. Caroline. 2005. *Mapping English Language Proficiency Scores Onto the Common European Framework*. TOEFL Research Report RR – 88. Princeton: ETS.

- Taylor, Lynda. 2005. 'Washback and Impact'. *ELT Journal* (59/2). 154-155.

- The Chauncey Group International Ltd. 1999. *TOEIC User Guide*. Princeton: ETS.

- The Chauncey Group International Ltd. date unknown. *TOEIC Technical Manual*. Princeton: ETS.

- Vollmer, Helmut. 2003. 'Leistungsmessung, Lernerfolgskontrolle und Selbstbeurteilung: Überblick'. In: Bausch, Karl-Richard; Christ, Herbert & Krumm, Hans-Jürgen (Eds.). *Handbuch Fremdsprachenunterricht*. Tübingen: A. Francke Verlag. 4th ed.

- Von der Hand, Gerhard. 2003. 'Qualitätssicherung und – entwicklung'. In: Bausch, Karl-Richard; Christ, Herbert & Krumm, Hans-Jürgen (Eds.): *Handbuch Fremdsprachenunterricht*. Tübingen: A. Francke Verlag. 4th ed.

- Wall, Dianne. 1997. 'Impact and Washback in Language Testing'. In: Clapham, C. & Corson, D. (Eds.).

- Weir, C.J. 1997. 'The Testing of Reading in a Second Language'. In: Clapham, C. & Corson, D. (Eds.).

- Wilson, Kenneth M. 1989. *Enhancing the Interpretation of a Norm-Referenced Second-Language Test Through Criterion Referencing: A Research Assessment of Experience in the TOEIC Testing Context*. RR 89-39. Princeton: ETS.

- Wilson, Kenneth M. 2000. *An Exploratory Dimensionality Assessment of the TOEIC Test*. Research Report RR 00-14. Princeton: ETS.

- Woodford, Protase E. 1982. 'The Test of English for International Communication (TOEIC)'. In: Brumfit, C.J. (Ed.): *English for International Communication*. Oxford: Pergamon Press.

- Beschluss der Kultusministerkonferenz vom 20.11.1998 in der Fassung vom 26.04.2002. *Rahmenvereinbarung über die Zertifizierung von Fremdsprachenkenntnissen in der beruflichen Bildung*. Beschlusssammlung der KMK. Beschluss Nr. 330.

- The Council of the European Union. *Application of the language rules at the Council*.
  http://www.consilium.europa.eu/cms3_fo/showPage.asp?lang=en&id=1255&mode=g&name=#
  (accessed 14/04/2008).

Appendix

N° 02981

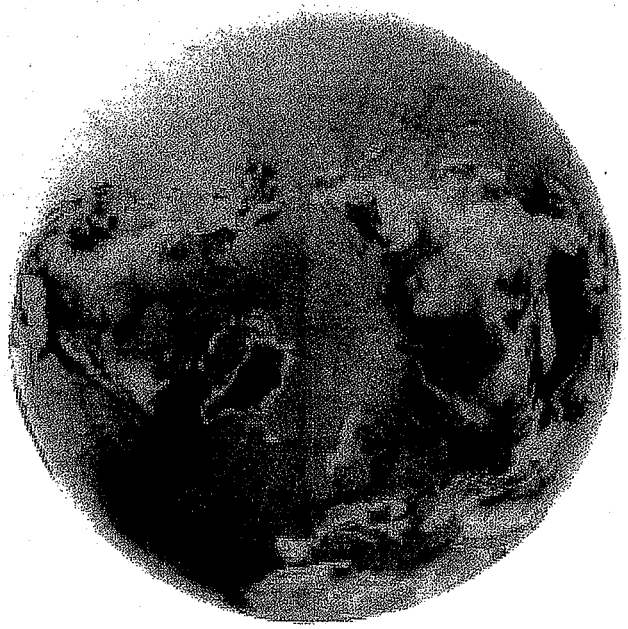
Registration Number Sample

Name Test

TOEIC Form ST-00

TOEIC®

TEST OF ENGLISH FOR INTERNATIONAL COMMUNICATION



Read the directions on the back cover.

Do not break the seal until you are told to do so.

This test book and the answer sheet must be handed in separately as instructed at the end of the test.

LISTENING COMPREHENSION

In this section of the test, you will have the chance to show how well you understand spoken English. There are four parts to this section, with special directions for each part.

PART I

Directions: For each question, you will see a picture in your test book and you will hear four short statements. The statements will be spoken just one time. They will not be printed in your test book, so you must listen carefully to understand what the speaker says.

When you hear the four statements, look at the picture in your test book and choose the statement that best describes what you see in the picture. Then, on your answer sheet, find the number of the question and mark your answer. Look at the sample below.

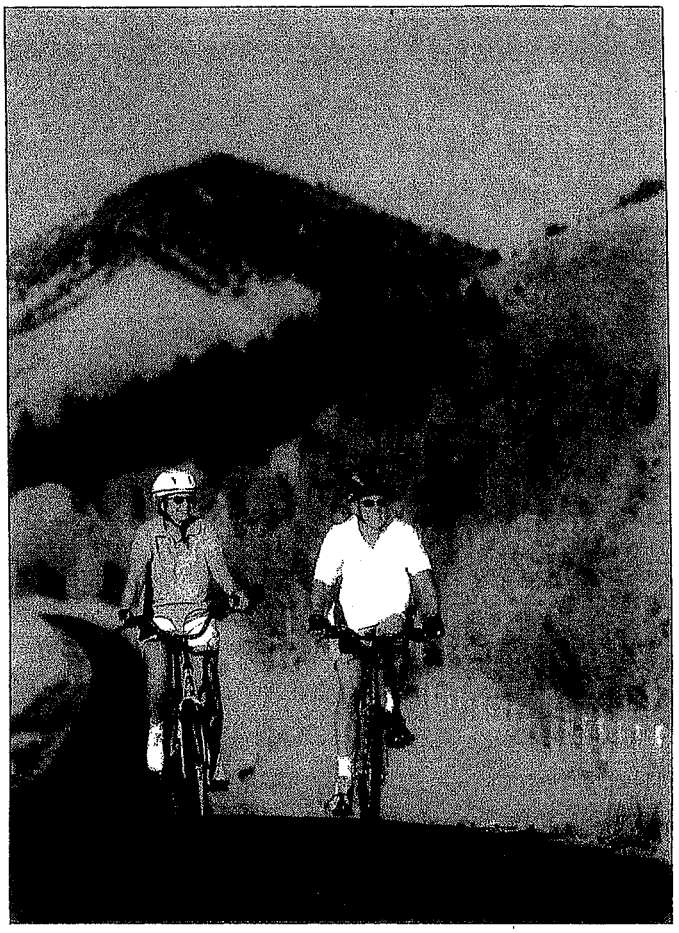
Sample Answer

Ⓐ ● Ⓒ Ⓓ

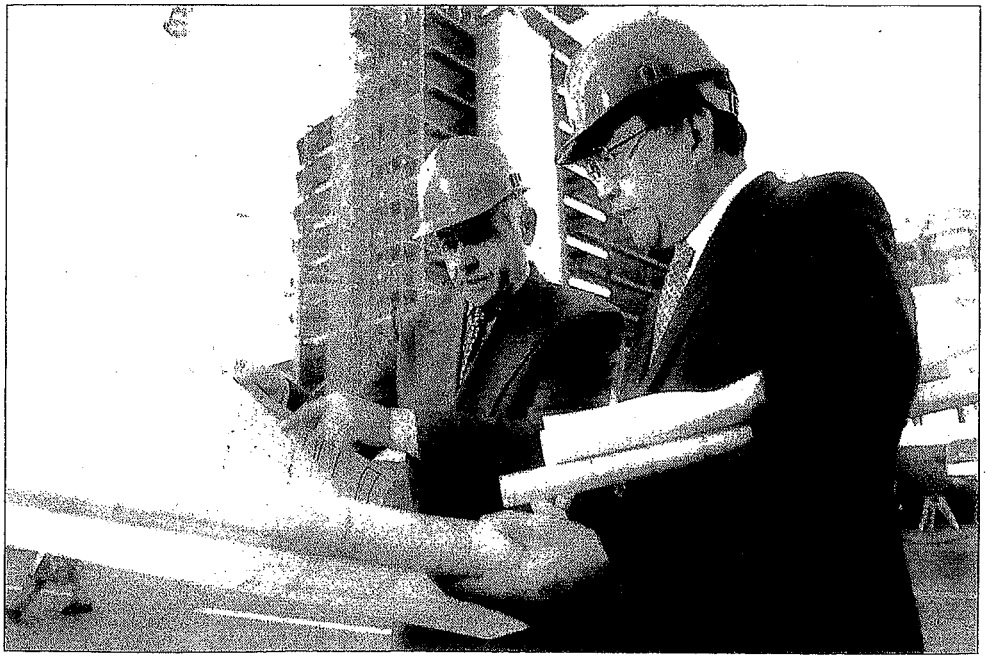
Now listen to the four statements.

Statement (B), "They're having a meeting," best describes what you see in the picture. Therefore, you should choose answer (B).

1.

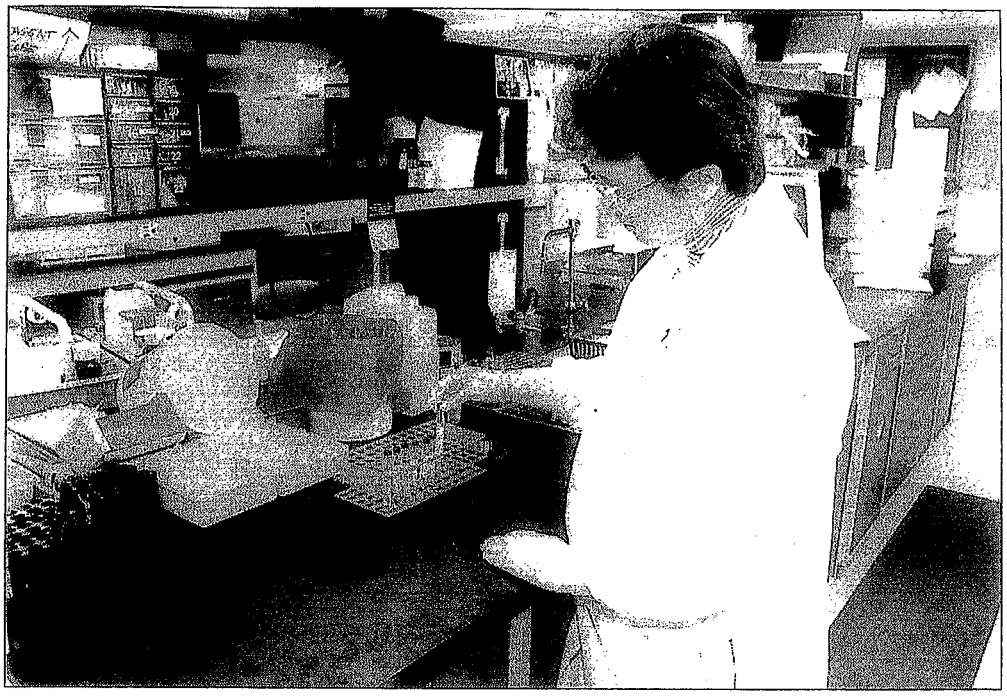


2.



GO ON TO THE NEXT PAGE

3.



4.

PART II

Directions: In this part of the test, you will hear a question or statement spoken in English, followed by three responses, also spoken in English. The question or statement and the responses will be spoken just one time. They will not be printed in your test book, so you must listen carefully to understand what the speakers say. You are to choose the best response to each question or statement.

Now listen to a sample question.

Sample Answer

You will hear:

● Ⓑ Ⓒ

You will also hear:

The best response to the question "How are you?" is choice (A), "I am fine, thank you." Therefore, you should choose answer (A).

21. Mark your answer on your answer sheet.

22. Mark your answer on your answer sheet.

23. Mark your answer on your answer sheet.

24. Mark your answer on your answer sheet.

25. Mark your answer on your answer sheet.

GO ON TO THE NEXT PAGE

PART III

Directions: In this part of the test, you will hear several short conversations between two people. The conversations will not be printed in your test book. You will hear the conversations only once, so you must listen carefully to understand what the speakers say.

In your test book, you will read a question about each conversation. The question will be followed by four answers. You are to choose the best answer to each question and mark it on your answer sheet.

51. Where do the speakers work?

(A) In a clothing store.
(B) At a newspaper office.
(C) At a law office.
(D) In a bookstore.

52. How do the speakers think the problem should be addressed?

(A) By admitting fewer patients.
(B) By hiring more doctors.
(C) By enlarging the waiting room.
(D) By changing the hours of operation.

53. When will the package arrive in London?

(A) Tuesday morning.
(B) Tuesday afternoon.
(C) Wednesday morning.
(D) Wednesday afternoon.

54. What is the man asking about?

(A) The cause of increased sales.
(B) Some overdue accounts.
(C) Reductions in sales staff.
(D) The results of a marketing survey.

55. Why will the speakers eat in the cafeteria today?

(A) They can meet colleagues there.
(B) The food is cheap there.
(C) The seafood is fresh there.
(D) They can eat quickly there.

PART IV

Directions: In this part of the test, you will hear several short talks. Each will be spoken just one time. They will not be printed in your test book, so you must listen carefully to understand and remember what is said.

In your test book, you will read two or more questions about each short talk. The questions will be followed by four answers. You are to choose the best answer to each question and mark it on your answer sheet.

81. Where is this talk most likely taking place?

(A) At a post office.
(B) At a factory.
(C) On an airplane.
(D) In a hospital.

82. What is being described?

(A) Flight schedules.
(B) Building repairs.
(C) Exit procedures.
(D) Work assignments.

83. In what area is Ms. Lee employed?

(A) Human resources.
(B) Telephone sales.
(C) Customer service.
(D) Product development.

84. What does Mr. Grieg want Ms. Lee to do?

(A) Update a map.
(B) Forward a list of names.
(C) Attend a board meeting.
(D) Accept a position.

This is the end of the Listening Comprehension portion of the test. Turn to Part V in your test book.

READING

In this section of the test, you will have the chance to show how well you understand written English. There are three parts to this section, with special directions for each part.

PART V

Directions: This part of the test has incomplete sentences. Four words or phrases, marked (A), (B), (C), (D), are given beneath each sentence. You are to choose the **one** word or phrase that best completes the sentence. Then, on your answer sheet, find the number of the question and mark your answer.

Example

Sample Answer
Ⓐ Ⓑ ● Ⓓ

Because the equipment is very delicate, it must be handled with -------.

(A) caring
(B) careful
(C) care
(D) carefully

The sentence should read, "Because the equipment is very delicate, it must be handled with care." Therefore, you should choose answer (C).

Now begin work on the questions.

101. Lyon Brothers, Inc., had a very small budget for advertising, so they decided to produce brochures -------.

(A) itself
(B) oneself
(C) ourselves
(D) themselves

102. Bianca Brunelli hopes to be ------- to government office in the spring.

(A) chosen
(B) elected
(C) preferred
(D) considered

103. City College is now offering programs designed for students ------- to pursue a two-year certificate in information technology.

(A) intending
(B) intended
(C) is intending
(D) has intended

104. All department supervisors are required to attend the ------- on the new employee time-keeping policy.

(A) delegation
(B) summary
(C) commission
(D) seminar

105. ------- the latest census, the population of the province has increased by eighteen percent in the last decade.

(A) In compliance with
(B) Depending on
(C) According to
(D) Along with

PART VI

Directions: In this part of the test, each sentence has four words or phrases underlined. The four underlined parts of the sentence are marked (A), (B), (C), (D). You are to identify the **one** underlined word or phrase that should be corrected or rewritten. Then, on your answer sheet, find the number of the question and mark your answer.

Example

Sample Answer

● Ⓑ Ⓒ Ⓓ

All employee (A) are required to wear (B) their identification (C) badges while (D) at work.

The underlined word "employee" is not correct in this sentence. This sentence should read, "All employees are required to wear their identification badges while at work." Therefore, you should choose answer (A).

Now begin work on the questions.

141. There are (A) several hotels in this area (B) that provides (C) discounts on tours of (D) historical sites.

142. Information collection (A) from shoppers through (B) surveys is stored in secure files and is used (C) to tailor direct mailings (D).

143. Even though (A) Ms. Herbert has been (B) director for six (C) months, she has not already visited (D) the branch offices.

144. For (A) personal reasons, Mr. Chun has decided not to (B) apply for a transference (C) at this time (D).

145. All household (A) chemicals they should (B) be stored well (C) out of the reach of children (D).

GO ON TO THE NEXT PAGE

PART VII

Directions: The questions in this part of the test are based on a selection of reading materials, such as notices, letters, forms, newspaper and magazine articles, and advertisements. You are to choose the **one** best answer, (A), (B), (C), or (D), to each question. Then, on your answer sheet, find the number of the question and mark your answer. Answer all questions following each reading selection on the basis of what is **stated** or **implied** in that selection.

Read the following example.

> The Museum of Technology is a "hands-on" museum, designed for people to experience science at work. Visitors are encouraged to use, test, and handle the objects on display. Special demonstrations are scheduled for the first and second Wednesdays of each month at 13:30. Open Tuesday-Friday 12:00-16:30, Saturday 10:00-17:30, and Sunday 11:00-16:30.

When during the month can visitors see special demonstrations?

Sample Answer

Ⓐ ● Ⓒ Ⓓ

(A) Every weekend
(B) The first two Wednesdays
(C) One afternoon a week
(D) Every other Wednesday

The reading selection says that the demonstrations are scheduled for the first and second Wednesdays of the month. Therefore, you should choose answer (B).

Now begin work on the questions.

Questions 163-164 refer to the following letter.

Kendar Office Supplies *Kemapriatarum Road, Bangkok 10110, Thailand*

Ms. Pranee Udomsak
Director
Beni & Beni, Inc.
426 Silom Road
Bangkok 10110
Thailand

Dear Ms. Udomsak:

In checking our records, I noticed that you are no longer listed as a current customer of Kendar Office Supplies. When I called and spoke to your office manager, Peri Davis, I was informed that your company is now using one of our competitors for your office needs. Ms. Davis referred me to you as the individual who makes all purchasing decisions at Beni & Beni.

Ms. Davis kindly described some of the problems that led you to select another supplier. I'm pleased to tell you that Kendar has made many improvements to its product line and services, and we are certain Beni & Beni will find these attractive. We have introduced a whole new line of office and computer supplies, many of which are not available from any other supplier. In addition, Kendar now has the largest warehouse facility in the region.

If you need any additional information please feel free to contact me. We welcome the opportunity to serve your company once again.

Sincerely,

Manee Chamchoy

Manee Chamchoy

163. For whom is this letter intended?

(A) The director of Beni & Beni
(B) The manager of Kendar Office Supplies
(C) Peri Davis
(D) Manee Chamchoy

164. What is the purpose of the letter?

(A) To verify customer data
(B) To register a formal complaint
(C) To inquire about warehouse space
(D) To restore a business relationship

GO ON TO THE NEXT PAGE

Questions 165-166 refer to the following information.

Electrical Safety Requirements and Procedures

An Up-to-Date, Intensive Two-Day Seminar

First Day

1. Introduction to Safety Standards
2. Conducting Electrical Inspections
3. Electrical Hazards
4. Training Requirements
5. Working on Energized Circuits or Parts
6. Installation of Electrical Equipment

Second Day

1. Personal Protection
2. Servicing of Electrical Equipment
3. Clearance Distance Guidelines
4. Electrical Hazards in Confined Spaces
5. Portable Electrical Equipment
6. Test Equipment
7. Protective Equipment

This course presents electrical safety information based on national industry regulations and is designed to meet and exceed national safety training for the field. We have no affiliation with any supplier or manufacturer. We are therefore able to present a completely neutral view of the industry, without the sales bias inherent in many supplier-sponsored programs. To generate free and open exchange of information, tape recording of course sessions will not be permitted.

For registration and fees call: (416) 555-1424 or visit our web site at www.taftt.com

The Association for Technological Training 3917 Stone St. TORONTO ON M5A 1N1

165. Which topic will be covered on the second day?

(A) Machinery installation guidelines
(B) Equipment maintenance and repair
(C) Hazardous waste disposal
(D) Personnel management techniques

166. What is a stated advantage of the seminar?

(A) The training is offered free of charge.
(B) Recordings of the sessions can be ordered.
(C) The course has no commercial sponsorship.
(D) Participants will receive training certificates.

Stop! This is the end of the test. If you finish before time is called, you may go back to Parts V, VI, and VII and check your work.

Correct Answers

Part I

1. C
2. D
3. A
4. B

Part II

21. B
22. A
23. B
24. C
25. A

Part III

51. B
52. C
53. C
54. A
55. D

Part IV

81. B
82. C
83. A
84. D

Part V

101. D
102. B
103. A
104. D
105. C

Part VI

141. C
142. A
143. D
144. C
145. B

Part VII

163. A
164. D
165. B
166. C

TOEIC®

Test of English for International Communication

General Directions

This is a test of your ability to use the English language. The total time for the test is approximately two hours. It is divided into seven parts. Each part of the test begins with a set of specific directions. Be sure you understand what you are to do before you begin work on a part.

You will find that some of the questions are harder than others, but you should try to answer each question to the best of your ability. Your score will be based on the number of questions you answer correctly.

Do not mark your answers in this test book. **You must put all of your answers on the separate answer sheet** that you have been given. When putting your answer to a question on your answer sheet, be sure to fill in the answer space corresponding to the letter of your choice. Fill in the space so that the letter inside the oval cannot be seen, as shown in the example below.

Mr. Jones ------- to his accountant yesterday.

Sample Answer

(A) talk
(B) talking
(C) talked
(D) to talk

The sentence should read, "Mr. Jones talked to his accountant yesterday." Therefore, you should choose answer (C). Notice how this has been done in the example given.

Mark only **one** answer for each question. If you change your mind about an answer after you have marked it on your answer sheet, completely erase your old answer and then mark your new answer. You must mark the answer sheet carefully so that the test-scoring machine can accurately record your test score.

ETS

TOEIC

MARKING DIRECTIONS

- Use only pencil
- Darken the circles completely
- Erase cleanly

CORRECT MARK ●

INCORRECT MARKS ⊘ ⊗ ◐ ⊙

1 NAME

At the arrow, using one box for each letter, print your family name(s) in the first area. If you have more than one family name, leave a space between names. In the second area, fill in your first name(s). Below each box, fill in the circle containing the same letter. If your name includes any accent marks, fill these in using the circles above the boxes.

2 SEX

○ M
○ F

3 COUNTRY CODE

4 LANGUAGE CODE

5 GROUP CODE (if assigned)

6 IDENTIFICATION NUMBER

7 DATE OF BIRTH

MM DD YYYY

8

JOB

ORGANIZATION

CITY

9

TESTING LOCATION

CITY

10

SIGNATURE DATE

73005-034502 • TF14E150 • Printed in U.S.A.
I.N.999973

Punkte

Aufgabe 2 **15**

Leseverstehen

Lesen Sie den Text auf der nächsten Seite durch und entscheiden Sie, ob nachfolgende Aussagen **richtig** oder **falsch** sind.

Nr.	Aussage	richtig	falsch
1.	Indien ist mittlerweile führend in der Entwicklung der Biotechnologie und alternativer Energien.		
2.	Die Verdienstmöglichkeiten für Spezialisten in Indien machen nur einen Bruchteil dessen aus, was sie in anderen Ländern verdienen.		
3.	Führungskräfte aus den westlichen Industrieländern werden von ihren eigenen Mitarbeitern kritisiert, da diese um ihre Arbeitsplätze bangen.		
4.	Verbindliche Vereinbarungen und Terminabsprachen werden zwar immer eingehalten, aber die indischen Angestellten äußern nicht immer offen ihre Meinung.		
5.	Frau Jhaveri profitiert von ihren Erfahrungen, da sie mit beiden Kulturen vertraut ist und glaubt, dass man nur so eine erfolgreiche Geschäftsbeziehung aufbauen kann.		
6.	Zwischen Deutschen und Indern bestehen häufig Kommunikationsprobleme, da die Inder sich sehr vage und indirekt äußern.		
7.	Die deutschen Geschäftsleute sind nicht an einer persönlichen Beziehung zu einem indischen Geschäftspartner interessiert.		
8.	Für indische Angestellte gibt es keine Probleme, die im Rahmen einer geschäftlichen Verbindung auftauchen können.		
9.	In indischen Familienbetrieben steht das persönliche Engagement und die Verantwortung gegenüber den Angestellten an erster Stelle.		
10.	Ein Managertyp in Indien verbindet die Tradition mit einem dynamischen Arbeitsethos.		

Punkte

Aufgabe 2

Anlage

A passage to India

According to many newspaper reports, India is Asia's „new tiger" and will soon be experiencing faster growth than China. The world's largest democracy has certainly made economic progress in recent years. As a result, "India Shining" was the name of the former government's recent optimistic election campaign.

Although serious problems remain, India is now a leader in IT services and call centres; it is also a major manufacturer of car parts, textiles and pharmaceuticals. It plays a significant role in the development of biotechnologies and alternative energy, and it is an increasingly important consumer market.

Highly qualified specialists who speak English and work for a fraction of what such specialists earn in the US or Western Europe, make India a popular location for offshore outsourcing. Its growth of around eight per cent a year has led many commentators to believe that India's time has finally come.

However, western managers involved in organizing projects in India sometimes see things less positively. They face the criticism from their own staff members at home, who fear that their jobs may be under threat. Bombarded with media images of extreme poverty and underdevelopment, workers in the West often think their Indian colleagues are less qualified than they are.

Another problem is that stereotypes of Indian business practices are confirmed when deadlines seem to be ignored and commitments are not met. Western managers complain that their Indian staff members don't take the initiative, don't give clear feedback, and never really show what they are thinking.

Trupti Jhaveri, an Indian CEO explains: "I am an Indian national who was born and raised in Germany. So I have been able to experience the benefits of both cultures. My involvement in intercultural coaching is a way of combining my academic interest in stereotypes with my personal and professional experience of both countries. I believe that understanding cultural backgrounds and social environments is a key factor in successful business interactions.

There are a number of differences in communication style. Germans often think that Indians are indirect and vague, and that it is difficult to get commitment from them. However, the Indians are interested in creating a positive atmosphere as the basis of long-term understanding. For Indians, the basis of trust is a relationship that covers the personal as well as the professional. Germans, on the other hand, often focus on momentary business interests and contractual commitments. Sometimes Germans don't want to "waste" time building relationships, as there is a high level of fluctuation in the Indian workforce and they think that will have to start all over again with the next person. The Indian, on the other hand, would like to be seen as a unique individual, not just as someone fulfilling a particular function."

It is difficult for Germans when an Indian says "no problem", as the German suspects that the real challenges are not being admitted. For the Indian "no problem" simply means: "I know there will be problems, but I'm doing the best I can at my end." Because Indians tend to think in terms of processes, they can be frustrated by the pressure to think in absolutes. They interpret this as poor problem-solving skills on the part of their German colleagues.

Management styles differ widely across cultures. In India it is important to differentiate between two generations of leadership. First, there are the family businesses. Here, the manager sets an example for the employees, and shows that his primary responsibility is to the company as a family, and not to his own interests. His personality and his role as a model are more important than formal qualifications.

The second type of management style is found in newer companies, and is influenced by the West. These managers blend traditional values with dynamic work ethics. They have a different sense of performance, are good at motivating, are highly trained and have international experience. Sometimes they can't see why they should stay in India – although, increasingly, non-resident Indians are returning to the country.

Punkte

Aufgabe 1 **15**

Hörverstehen

Teil 1

Zwei persönliche Assistenten, Sandra aus Deutschland und Gary aus den USA, vergleichen ihre Arbeit.

Hören Sie aufmerksam zu und entscheiden Sie, ob die folgenden Aussagen **richtig** oder **falsch** sind. Kreuzen Sie dazu bitte die entsprechenden Felder in der Tabelle an. Sie hören den Text zwei Mal.

Nr.	Aussage	richtig	falsch
1.	Sandra muss nicht immer bis 20 Uhr arbeiten, da sie sich mit einer Kollegin abwechseln kann.		
2.	Gary muss täglich über 200 E-Mails beantworten.		
3.	Sandra bekommt wenige E-Mails von außerhalb ihres Unternehmens.		
4.	Gary muss für seinen Chef Roger Vorträge schreiben, die dieser auf verschiedenen Sitzungen hält.		
5.	Sandra begleitet ihren Chef regelmäßig zu seinen Sitzungen.		
6.	Aufgrund des intensiven Kontaktes mit ihrem Chef darf Sandra ihn beim Vornamen nennen.		
7.	Reisevorbereitungen trifft Gary für seinen Chef ausschließlich online.		
8.	Um die Erstattung der Reisekosten kümmert sich Roger selbst.		
9.	Das Unternehmen, in dem Sandra arbeitet, wurde mit einem amerikanischen Unternehmen fusioniert.		

Teil 2

Sie hören Ihren Anrufbeantworter ab, auf dem eine Nachricht hinterlassen wurde. Hören Sie aufmerksam zu und tragen Sie die Angaben in die entsprechenden Kästchen ein. Sie hören die Nachricht zwei Mal.

10. Name:																			
11. Order No:																			
12. Phone No:																			

Punkte

Aufgabe 3 **30**

Schriftstücke erstellen

Sie schreiben an die Firma Schön & Co. GmbH Büroausstattungen, Am Sportplatz 23, 68237 Mannheim eine Einladung zur Messe „Office of the future“ in Hartford, Connecticut, USA.

Schreiben Sie die Einladung in englischer Sprache. Ihre Aufgabe ist es lediglich, einen Brieftext zu verfassen unter Berücksichtigung der erforderlichen Formvorschriften und des folgenden Inhalts:

- Termin: 17.-22. August 2006; Ort: Hartford; Thema: siehe oben
- Bieten Sie einen Stand von 35m² an, der eine Rückwand, zwei Seitenwände sowie eine abschließbare Kabine hat.
- Der Preis für den Stand liegt bei $450 pro Tag.
- Es besteht die Möglichkeit, ihren Kunden am Stand Tee und Kaffee anzubieten.
- Außerdem gibt es Regale, auf denen Verkaufsunterlagen und Faltblätter ausgelegt werden können.
- Darüber hinaus gibt es genug Platz, um einige der Kopiergeräte auszustellen.
- Bieten Sie an, bei der Buchung einer Unterkunft behilflich zu sein.
- Das Messegelände ist stündlich mit dem Bus von der Stadtmitte aus zu erreichen, was Ihnen abends die Gelegenheit bietet, Hartford kennen zu lernen.
- Es werden über 200 Hersteller auf dieser Messe ausstellen und man erwartet über 500.000 Besucher.
- Abschlusssatz, in dem Sie zum Ausdruck bringen, dass Sie hoffen, Interesse geweckt zu haben und auf baldige Antwort warten.

Punkte

Aufgabe 3

Antwortseite (benutzen Sie ggf. ein zusätzliches leeres Blatt)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Punkte

Aufgabe 4 **10**

Schriftstücke erstellen

Jane Law ist Gebietsmanagerin von Modern Clothing, einem Bekleidungsgeschäft, für das Sie arbeiten. Sie ist besorgt über die schlechte Geschäftslage Ihrer Filiale. Eine Unternehmensberatungsgesellschaft wird eingeschaltet, deren Tätigkeit in einer E-Mail angekündigt werden soll.

Verwenden Sie dazu die nachstehenden Stichwortvorgaben, die sinngemäß ins Englische zu übertragen sind:

- Wie angekündigt im letzten Rundbrief, wird die Unternehmensberatungsgesellschaft McPINSEY vom 27.03. – 07.04.06 in der Filiale tätig werden.
- McPINSEY hat in den letzten 12 Jahren Unternehmen in vielen Projekten erfolgreich beraten.
- McPINSEYs konsequentes Qualitätsmanagement, ständige Weiterbildung, teamorientiertes Arbeiten und internationale Verbindungen garantieren optimale Lösungen.
- Stärke liegt im Fachwissen in den Bereichen Datenverarbeitung und Organisationswesen und dem „Ohr am Markt".
- Vorschläge von Mitarbeitern zur Verbesserung Lage des Unternehmens willkommen.

Benutzen Sie das beigefügte Formular und verwenden Sie gegebenenfalls ein weiteres Blatt.

e-mail ☒

Mail from:	
Mail to:	
Cc:	
Subject:	
Date:	May 11, 2006
Time:	09:00
Attachment:	

.....

.....

.....

.....

.....

.....

Punkte

Aufgabe 4

Antwortseite (benutzen Sie ggf. ein zusätzliches leeres Blatt)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Punkte

Aufgabe 5 **30**

Mediation

Ihr Vorgesetzter bittet Sie, folgende Beschreibungen der Dienstleistungen eines Büroausstatters ins Englische zu übertragen:

- Preisgarantie: wenn Kunden Produkt zu günstigerem Ladenpreis finden, Erstattung von 110% der Preisdifferenz innerhalb von 2 Wochen nach Kauf
- - Voraussetzung: gleiche Artikelnummer und vorrätig bei anderem Einzelhändler
- - Ausschluss von Internet- und Eröffnungsangeboten sowie Geschäftsauflösungen
- Lieferbedingungen: nach Auftragserteilung Erhalt eines voraussichtlichen Liefertermins
- - bei Bestellung bis 17.00 Uhr Ortszeit und Verfügbarkeit des Produkts, Versand von Auftragsbestätigung am gleichen Tag
- - Für Gewährleistung des sicheren Erhalts der Ware Erreichbarkeit zwischen 9:00 und 17:00 Uhr unerlässlich zwecks Bestätigung der Lieferung per Unterschrift
- - Bei Nichtantreffen des Kunden, Benachrichtigung mit Anweisung bezüglich Abholungsmodalitäten
- Ersatzlieferung bis ein Jahr nach Ablauf der Gewährleistungsfrist; einzigartige Garantiefrist in dieser Branche
- Online Reparaturservice bei technischem Versagen an sieben Tagen rund um die Uhr, Kontakt mit der technischen Service-Hotline über Webdialog oder Verweis an Vertragshändler
- Bei Kauf von Büromöbeln Versicherung gegen Verschleißschäden möglich, jährliche Versicherungsprämie $65

Punkte

Aufgabe 5

Antwortseite (Benutzen Sie ggf. ein zusätzliches leeres Blatt)

- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....
- .....
  .....
  .....

Kaufm.-verwaltende Berufe **Best-practice** KMK Stufe II (B1)

Sie sind beauftragt, gemeinsam mit einem englischsprachigen Kollegen ein Kundeninformationsbüro einzurichten. Informieren Sie Ihren Kollegen über:

- Lage (Erdgeschoss, Eingangsbereich, etc.)
- Größe (80 qm, Fensterfront zur Straße, etc.)
- Funktion des Raumes (Kundenbetreuung)

Diskutieren Sie die Ausstattung/Einrichtung des Büros. Für Sie ist besonders die kundenfreundliche Atmosphäre wichtig, wie zum Beispiel:

- Sitzgruppe
- Pflanzen
- Platz für Informationsmaterialien

Einigen Sie sich angesichts des knappen Budgets auf die wichtigsten Anschaffungen.

Sie sind beauftragt, gemeinsam mit einem englischsprachigen Kollegen ein Kundeninformationsbüro einzurichten. Informieren Sie Ihren Kollegen über:

- das Budget (20.000,00 €)
- Fertigstellung des Büros: Ende des Jahres
- Technische Ausstattung von drei Arbeitsplätzen

Diskutieren Sie die Ausstattung/Einrichtung des Büros. Die technische Ausstattung sollte Ihrer Meinung nach optimal sein:

- IT Ausstattung
- Farb-Laser-Kopierer
- Touch-Screen für Kundeninformation

Einigen Sie sich angesichts des knappen Budgets auf die wichtigsten Anschaffungen.

Fragebogen
Verbreitung externer Sprachtests
an weiterführenden Schulen

Als Lehramtsstudent für Englisch, Philosophie/Ethik und Deutsch möchte ich in meiner Zulassungsarbeit (wissenschaftlichen Arbeit) für das erste Staatsexamen ermitteln, in wieweit bereits externe Sprachtests an weiterführenden Schulen eingeführt sind und welche Überlegungen es generell zu diesen gibt.

1) Name der Schule: _____

2) Schulform: _____

3) Befindet sich die Schule im ☐ ländlichen oder ☐ städtischen Raum?

4) Wie viele Schüler besuchen Ihre Schule? _____

5) Wie groß ist der Abschlussjahrgang? _____

6) Gibt es einen sprachlichen Zug? ☐ ja ☐ nein

7) Ist die Förderung von Fremdsprachen Teil des Schulprofils? ☐ ja ☐ nein

8) Welche Sprachen werden angeboten?

	regulär	nur als AG
Englisch	☐	
Französisch	☐	
Latein	☐	☐
Spanisch	☐	☐
Italienisch	☐	☐
Russisch	☐	☐
Chinesisch	☐	☐
Portugiesisch	☐	☐
Türkisch	☐	☐
andere:	☐	☐
andere:	☐	☐

9) Bieten Sie externe Sprachtests an?

Englisch	☐ nein weiter mit 10)	☐ ja, den weiter mit 12)	
Französisch	☐ nein	☐ ja, den	_____
Spanisch	☐ nein	☐ ja, den	_____
Italienisch	☐ nein	☐ ja, den	_____
			_____

Bei den weiteren Fragen konzentrieren wir uns auf die englischen Sprachtests

10) Wir bieten momentan keine Sprachtests, da … (Mehrfachnennungen möglich)

☐ kein Bedarf
☐ zu teuer
☐ zu aufwändig in der Durchführung
☐ zu aufwändig in der Vorbereitung der Schüler auf den Test
☐ noch gar nicht daran gedacht
☐ wird gerade überlegt
☐ _____

11) Unter welchen Voraussetzungen können Sie sich vorstellen, Sprachtests an der Schule anzubieten? (Mehrfachnennungen möglich)

☐ wissenschaftliche Fundierung des Testaufbaus
☐ Anerkennung auf dem internationalen Markt

☐ günstiger Preis, max. _____ €
☐ Berücksichtigung des gemeinsamen europäischen Referenzrahmens (CEF)

☐ geringer Aufwand in der Verwaltung, _____ Stunden max.

☐ geringer Aufwand in der Vorbereitung der Schüler auf den Test, max. _____ Stunden
☐ Nachfrage von Schüler-/Elternseite aus
☐ als Immatrikulationsbedingung bei Hochschulen erforderlich

☐ _____

weiter mit Frage 14

12) Wir haben uns aus folgenden Gründen für den oben genannten Sprachtest entschieden (Mehrfachnennungen möglich):

☐ Bekanntheitsgrad
☐ Anerkennung auf dem internationalen Markt
☐ eigene Erfahrung mit dem Test
☐ Qualität
☐ auf Anregung von Schülern/Eltern
☐ im Rahmen der Profilbildung der Schule
☐ auf Anregung der Schulbehörde

☐ _____

13) Unter welchen Umständen wären Sie bereit, einen anderen Sprachtest einzuführen (Mehrfachnennung möglich):

☐ geringerer Preis, kleiner als _____ €
☐ geringerer Aufwand in der Verwaltung, weniger als _____ Stunden
☐ geringerer Aufwand in der Vorbereitung der Schüler auf den Test, weniger als _____ Stunden
☐ höhere Anerkennung auf dem internationalen Markt
☐ leichter zu interpretierendes Ergebnis
☐ als Immatrikulationsbedingung bei Hochschulen erforderlich
☐ Wir sind vollauf zufrieden. Ein Wechsel ist uninteressant.

14) Welche der folgenden englischen Sprachtests kennen Sie?

☐ KMK-Zertifikat	☐ Cambridge (FCE, CAE …)	☐ TOEIC
☐ TELC	☐ Trinity	☐ LCCI
☐ TOEFL	☐ IELTS	☐ BEC

Vielen Dank für Ihre Teilnahme.

Bitte senden Sie diesen Fragebogen per E-Mail an sebastian.kluitmann@mars.uni-freiburg.de oder per Fax an 0711/45 10 17-377

Falls Sie den Postweg bevorzugen:
Sprachenmarkt.de
z.Hd. Sebastian Kluitmann
Wollgrasweg 49
70599 Stuttgart

Für Rückfragen erreichen Sie mich unter 0711/45 10 17-373

Wenn Sie am Ergebnis des Fragebogens interessiert sind, nennen Sie mir bitte Ihre E-Mail-Adresse:

Bei der Erstellung meiner Studie werde ich im Rahmen eines Praktikums von Sprachenmarkt.de, einem Unternehmen, das schulbezogene Dienstleistungen wie Lehrerfortbildungen /-kurse im Rahmen des COMENIUS 2.2.c Programms, Organisation und Durchführung von Sprachtests an Ihrer Schule, Klassenfahrten, Sprachkurse, etc. anbietet, unterstützt.

Bitte besuchen Sie die Internetseite www.sprachenmarkt.de oder wenden Sie sich telefonisch unter 0711/45 10 17-370 an uns, wenn Sie weitere Informationen wünschen.

kannt. Dieses Bekenntnis muss in den kommenden Jahren durch verbesserte Rahmenbedingungen und eine angemessene Lehrerversorgung umgesetzt werden und darf nicht nur ein Lippenbekenntnis bleiben.

Zur Qualitätssicherung unserer Berufsschule haben wir auch in diesem Jahr an unserem Leitbild gearbeitet (siehe Bericht). Wir setzten uns Ziele und haben auch den Mut uns daran messen zu lassen.

In seiner Weihnachtsansprache wies Bundespräsident Horst Köhler auf die Bedeutung unserer gemeinsamen Anstrengungen hin, indem er betonte: „Arbeit hilft uns, das Leben aus eigener Kraft zu meistern und vermittelt Lebenssinn. Deshalb müssen wir alles tun, damit die jungen Menschen Zugang zum Berufsleben finden.“.

Ich wünsche Ihnen eine erfolgreiche Zeit und schöne Fastnachtstage.

Axel Rombach, Abteilungsleiter KBS

KMK – Zertifikat an unserer Schule

Im Schuljahr 1999/2001 wurden in Baden-Württemberg erstmals berufsbezogene Fremdsprachenkenntnisse, die heutzutage von Schülern und Auszubildenden als Teil des Fachwissens und beruflicher Fertigkeiten erwartet werden, im Rahmen eines Pilotversuchs an verschiedenen kaufmännischen Berufsschulen zertifiziert.

Die Vermittlung berufsbezogener, fremdsprachlicher Qualifikationen ist im Wahlpflichtbereich als Erweiterungsunterricht mit bis zu 80 Stunden (2 Stunden pro Woche) vorgesehen und dient der Vorbereitung auf das KMK-Fremdsprachenzertifikat.

Dieses Zertifikat, das auf die Europarat-Initiative „Europäischer Referenzrahmen für das Lernen und Lehren von Sprachen“ zurückgeht, ist europaweit vergleichbar.

Berufliche Schulen können auf freiwilliger Basis - unabhängig von einer Benotung im Zeugnis - eine Prüfung anbieten, in der sich Schülerinnen und Schüler ihre Fremdsprachenkenntnisse zertifizieren lassen können. Die jeweilige Schule entscheidet, ob und für welche Schüler eine Prüfung nach diesen Richtlinien durchgeführt wird.

Die Prüfung besteht aus einem schriftlichen und einem mündlichen Teil. Es werden die folgenden Kompetenzbereiche zu Grunde gelegt:

Rezeption (Fähigkeit, gesprochene und geschriebene fremdsprachliche Mitteilungen zu verstehen)

Produktion (Fähigkeit, sich mündlich und schriftlich in der Fremdsprache zu äußern)

Mediation (Fähigkeit, durch Übersetzung oder Umschreibung mündlich oder schriftlich zwischen Kommunikationspartnern zu vermitteln).

Interaktion (Fähigkeit, Gespräche zu führen und zu korrespondieren)

Die Zertifikatsprüfungen werden in Anlehnung an die Lehrpläne der Berufsschule erstellt und finden an den Schulen statt. Das Zertifikat wird von der Schule vergeben. Das den erfolgreichen Teilnehmern zu verleihende Zertifikat enthält auf der Vorderseite keine Note, sondern detaillierte Angaben über die Prüfungsteile sowie die in den einzelnen Kompetenzbereichen erzielten Ergebnisse.

Wir werden die Prüfung zu diesem KMK-Zertifikat am 16. März 2007 für die Industriekaufleute und am 8. Mai 2007 für die IT-Systemkaufleute anbieten.

len und ein Investitionsprogramm für den Ausbau von Ganztagesschulen aufzulegen, bei dem insbesondere jene Schulträger zum Zuge kommen, die aufgrund der Überzeichnung des Bundesprogramms IZBB keine Investitionsmittel mehr erhalten;

Hauptschulen mit besonderer pädagogischer und sozialer Aufgabenstellung sind im Wege des Schulversuchs als Ganztagesschulen eingerichtet. Daneben gibt es im Land Ganztagesschulen als Angebote auf freiwilliger Grundlage. Schülerinnen und Schüler, die sich in diesen Schulen angemeldet haben, besuchen die Schule mit allen Rechten und Pflichten, sodass auch die nachmittägliche Anwesenheit zur Schulpflicht dazu gehört.

Im Übrigen kommt es in allen Schularten vor, dass aus stundenplantechnischen Gründen an einzelnen Nachmittagen Unterricht stattfindet.

Eltern bestimmen in eigener Verantwortung, ob die Erziehung und Betreuung ihres Kindes innerhalb oder außerhalb der eigenen Familie erfolgen soll. Ganztagsangebote an Schulen können daher nur ein freiwilliges Angebot für Eltern und deren Kinder sein. Die Landesregierung beabsichtigt daher derzeit nicht, Ganztagsschulen im Schulgesetz zu verankern. Dies wäre vor allem dann notwendig, wenn sie für die Schülerinnen und Schüler zur Pflicht werden sollte. Es wird aber in den kommenden Jahren vor allem darum gehen, Ganztagesschulen als Angebote auszubauen.

Das Land unterstützt den Ganztagesbetrieb an Hauptschulen und einzelnen Grundschulen, die als Schulen mit besonderer pädagogischer und sozialer Aufgabenstellung (sog. Brennpunktschulen) eingestuft sind, sowie an Förderschulen in enger räumlicher Nähe zu einer Brennpunkthauptschule mit einer zusätzlichen Lehrerzuweisung. Im Hinblick auf die Haushaltslage des Landes ist die Finanzierung von pädagogischem Personal an anderen Schulen und Schularten derzeit nicht leistbar.

In seiner Regierungserklärung am 27. April 2005 im Landtag hat Ministerpräsident Oettinger angekündigt, neben dem Ausbau von Brennpunktschulen in Ganztagsform in den kommenden Jahren bedarfsorientiert neue Ganztagsschulen in offener Angebotsform in allen Schularten einzurichten und qualifizierte Jugendbegleiter in der Ganztagsbetreuung einzusetzen. In einem Gespräch zwischen dem Land und den Kommunalen Landesverbänden wurde vereinbart, eine Förderung der Ganztagesschulangebote sowie die Abwicklung unerledigter Zuschussanträge aus dem Investitionsprogramm des Bundes „Zukunft Bildung und Betreuung“ (IZBB) aus Mitteln des Kommunalen Investitionsfonds (KIF) zu prüfen. In diesem Zusammenhang wird eine Neufassung der Schulbauförderrichtlinien mit dem Ziel einer generellen Förderung von Ganztagesangeboten erfolgen.

6. Berufliches Schulwesen

das berufliche Schulwesen einschließlich der beruflichen Gymnasien und Berufskollegs auszubauen und insbesondere den strukturellen Unterrichtsausfall abzubauen;

Bereich Berufsschule

Kernpunkt der Anstrengungen der Landesregierung zur Weiterentwicklung im Bereich der Berufsschule bildet die zeitnahe Umsetzung der auf Bundesebene neu geschaffenen bzw. neu geordneten Berufsbilder, die auf den Erhalt respektive eine Stärkung der Ausbildungsbereitschaft der Wirtschaft zielt. Durch die enge Verflechtung mit den wirtschaftlichen und technologischen

Entwicklungen ist der Bereich der Berufsschule, wie keine andere Schulart, nicht nur von einem andauernd hochdynamischen Innovationsprozess geprägt, sondern auch direkt von der konjunkturellen Lage abhängig. Dennoch ist es durch die gemeinsame Anstrengung aller Beteiligten im Rahmen des Bündnisses zur Stärkung der beruflichen Ausbildung in Baden-Württemberg im letzten Jahr gelungen, den Abwärtstrend der Vorjahre umzukehren. So stieg die Zahl der im Land neu abgeschlossenen Ausbildungsverträge in 2004 im Vergleich zum Vorjahr um 3,5 Prozent und liegt damit über dem Bundesdurchschnitt von 2,8 Prozent.

Um die Berufsausbildung im dualen System auch in Zukunft attraktiv zu halten, sind in enger Kooperation mit der Wirtschaft zahlreiche zertifizierbare Zusatzqualifikationen entstanden, wie z. B. berufsbezogenes Englisch mit Europazertifikat (KMK-Zertifikat), Qualitätsmanagement und Kundenservice sowie vieles mehr, die nachfrageorientiert entsprechend den Bedürfnissen der Ausbildungsbetriebe vor Ort ausgebaut und weiterentwickelt werden. Zusätzlich besteht seit mehreren Jahren die Möglichkeit, parallel zur Berufsausbildung die Fachhochschulreife zu erwerben, wovon von den Schülerinnen und Schülern in zunehmendem Maße Gebrauch gemacht wird.

Bereich Berufskolleg

Die Landesregierung verfolgt seit Jahren das Ziel, die Zahl der Absolventinnen und Absolventen mit Hochschulzugangsberechtigung (Fachhochschulreife, Hochschulreife) zu steigern. Deshalb hat sie, um die Durchlässigkeit der Berufskollegs weiter zu verbessern, in die Stundentafeln aller Berufskollegs ein Zusatzprogramm zur Erlangung der Fachhochschulreife aufgenommen. Damit und durch die Einrichtung von weiteren Klassen in den Berufskollegs konnte die Zahl der Absolventinnen und Absolventen, die mit Fachhochschulreife die öffentlichen und privaten Schulen im Jahr 2004 verlassen haben, auf insgesamt 13.438 gesteigert werden.

Bereich berufliche Gymnasien

Die Weiterentwicklung der beruflichen Gymnasien mit der Stärkung der Kernkompetenzfächer im Allgemeinen und dem Ausbau der spezifischen Profile bzw. Richtungen – vor allem im Bereich der Zukunftstechnologien – im Besonderen soll auch den Anforderungen des Wirtschaftsstandortes Baden-Württemberg hinsichtlich der Sicherung eines qualifizierten Nachwuchses Rechnung tragen. Ziel der Weiterentwicklung ist insbesondere ein stärkeres Lernen von Grundlagen und die Sicherstellung der Studierfähigkeit. Die erste Abiturprüfung nach der neuen Verordnung über Abitur und Versetzung an beruflichen Gymnasien (BGVO) fand im Frühjahr 2005 statt. In einem stetigen Weiterentwicklungsprozess wurde der Schulversuch „Berufliches Gymnasium der biotechnologischen Richtung“ in die Regelform überführt. Nach dem im Schuljahr 2004/05 erreichten Ausbaustand wird das biotechnologische Gymnasium inzwischen an 24 Standorten angeboten. Beginnend mit dem Schuljahr 2005/06 wird – zunächst an vier Standorten – das neue Profil „Technik und Management“ an Technischen Gymnasien angeboten. Das neue Profil umfasst technische Inhaltsschwerpunkte und verknüpft sie mit vertieften wirtschaftswissenschaftlichen Kenntnissen. Dort wird der Gedanke der interdisziplinären Anlage von Bildungsgängen aufgegriffen, wie er beispielsweise in den in Karlsruhe und Stuttgart angebotenen Studiengängen des Wirtschaftsingenieurs (Universität/Fachhochschule) erfolgreich praktiziert und in der Arbeitswelt zunehmend gefordert wird.

Die Zahl der Schülerinnen und Schüler steigt an den beruflichen Gymnasien stetig an. Die Schülerzahlen an den öffentlichen beruflichen Gymnasien sind

Principles and Practice in Language Testing

Compliance or Conflict?

Or

European Standards in Language Assessment?

Outline

- The Past
- The Past becoming Present – Present Perfect?
- The Future?

Standards?

Shorter OED:

- Standard of comparison or judgement
- Definite level of excellence or attainment
- A degree of quality
- Recognised degree of proficiency
- Authoritative exemplar of perfection
- The measure of what is adequate for a purpose
- A principle of honesty and integrity

Standards?

- Report of the Testing Standards Task Force,

  ILTA 1995 (International Language Testing Association = ILTA)

  http://www.iltaonline.com/ILTA_pubs.htm

1. Levels to be achieved
2. Principles to follow

Standards as Levels

- FSI/ILR/ACTFL/ASLPR
- Foreign Service Institute
- Interagency Language Round Table
- American Council for the Teaching of Foreign Languages
- Australian Second Language Proficiency Ratings

Standards as Levels

Europe?

- Beginner/ False Beginner/Intermediate/Post Intermediate/Advanced
- How defined?
- Threshold Level?

Standards as Principles

- Validity
- Reliability

- Authenticity?
- Washback?
- Practicality?

Standards as Principles

The psychometric tradition

- Tests externally developed and administered
- National or regional agencies responsible for development, following accepted standards
- Tests centrally constructed, piloted and revised
- Difficulty levels empirically determined
- Externally trained assessors
- Empirical equating to known standards or levels of proficiency

Standards as Principles

In Europe:

- Teacher knows best
- Having a degree in a language means you are an ‘Expert’
- Experience is all
- But 20 years experience may be one year repeated twenty times! and is never checked

Past (?) European tradition

- Quality of important examinations not monitored
- No obligation to show that exams are relevant, fair, unbiased, reliable, and measure relevant skills
- University degree in a foreign language qualifies one to examine language competence, despite lack of training in language testing
- In many circumstances merely being a native speaker qualifies one to assess language competence.
- Teachers assess students' ability without having been trained.

Past (?) European tradition

- Teacher-centred
- Teacher develops the questions
- Teacher's opinion the only one that counts
- Teacher-examiners are not standardised
- Assumption that by virtue of being a teacher, and having taught the student being examined, teacher-examiner makes reliable and valid judgements
- Authority, professionalism, reliability and validity of teacher rarely questioned
- Rare for students to fail

Past becoming Present: Levels

- Threshold 1975/ Threshold 1990
- Waystage/ Vantage
- Breakthrough/ Effective Operational / Mastery
- CEFR 2001
- A1 – C2
- Translated into 23 languages so far, including Japanese!

Past becoming Present: Levels

- CEFR enormous influence since 2001
- ELP contributes to spread
- Claims abound
- Not just exams but also curricula/ textbooks

Manual for linking exams to CEFR

- Familiarisation – essential, even for "experts" – Knowledge is usually superficial
- Specification
- Standard setting
- Empirical validation

Manual for linking exams to CEFR

BUT FIRST

- If an exam is not valid or reliable, it is meaningless to link it to the CEFR

How can validity be established?

- My parents think the test looks good.
- The test measures what I have been taught.
- My teachers tell me that the test is communicative and authentic.
- If I take the Abitur instead of the FCE, I will get the same result.
- I got a good English test result, and I had no difficulty studying in English at university.

How can validity be established?

- Does the test match the curriculum, or its specifications?
- Is the test based adequately on a relevant and acceptable theory?
- Does the test yield results similar to those from a test known to be valid for the same audience and purpose?
- Does the test predict a learner's future achievements?

How can validity be established?

Note: a test that is not reliable cannot, by definition, be valid

- All tests should be piloted, and the results analysed to see if the test performed as predicted
- A test's items should work well: they should be of suitable difficulty, and good students should get them right, whilst weak students are expected to get them wrong.

Factors affecting validity

- Lack of specifications
- Lack of training of item/ test writers
- Lack of / unclear criteria for marking
- Lack of piloting/ pre-testing
- Lack of detailed analysis of items/ tasks
- Lack of standard setting
- Lack of feedback to candidates and teachers

Standards as Principles: Reliability

- If I take the test again tomorrow, will I get the same result?
- If I take a different version of the test, will I get the same result?
- If the test had had different items, would I have got the same result?
- Do all markers agree on the mark I got?
- If the same marker marks my test paper again tomorrow, will I get the same result?

Factors affecting reliability

- Poor administration conditions – noise, lighting, cheating
- Lack of information beforehand
- Lack of specifications
- Lack of marker training
- Lack of standardisation
- Lack of monitoring

Present: Practice and Principles

- Teacher-based assessment vs central development
- Internal vs external assessment
- Quality control of exams or no quality control
- Piloting or not
- Test analysis and the role of the expert
- The existence of test specifications – or not
- Guidance and training for test developers and markers – or not

Present Perfect?

Exam Reform in Europe (mainly school-leaving exams)

- Slovenia
- The Baltic States
- Hungary
- Russia
- Slovakia
- Czech Republic
- Poland
- Germany

Present Perfect: Positive features

- National exams, designed, administered and marked centrally
- External exam replaces locally produced, poor quality exams
- National and regional exam centres to manage the logistics
- Results are comparable across schools and provinces
- Exams are recognised for university entrance

Present Perfect: Positive features

- Tests of communicative skills rather than traditional grammar
- Teams of testing experts firmly located in classrooms have been developed
- Items developed by teams of trained item writers
- Tests piloted and results analysed
- Rating scales developed for rating performances
- Scripts anonymised and marked by trained examiners, not own class teacher

Present Perfect: Positive features

- Secondary school teachers are involved in all stages of test development
- Nature and rationale for changes communicated to teachers
- Many training courses for teachers, including explicit guidance on exam preparation
- Teachers largely enthusiastic about the changes
- Positive washback claimed by teachers

Present Perfect: Positive features

- Exams beginning to be related to CEFR
- Comparability across cities, provinces, countries and regions
- Transparency, recognition and portability of qualifications
- Valuable for employers
- Yardstick for evaluating achievement of pupils and schools

Unprofessional

- No piloting, especially of Speaking and Writing tasks
- Leaving speaking tasks up to teachers to design and administer, typically without any training in task design
- Administering speaking tasks to Year 9 students in front of the whole class
- Administering speaking tasks to one candidate whilst four or more others are preparing their performance in the same room

Unprofessional

- No training of markers
- No double marking
- No monitoring of marking
- No comparability of results across schools, across markers/towns/ regions or across years (test equating)
- No guidance on how to use centrally devised scales, how to resolve differences, how to weight different components, no guidance on what is an “adequate” performance

Unprofessional

- No developed item production system:
- Pre-setting cut scores without knowledge of test difficulty
- No understanding that the difficulty of a task item or test will affect the appropriacy of a given cut-score
- Belief that a “good teacher” can write good test items: that training, moderation, revision, discussion, is not needed
- Lack of provision of feedback to item writers on how their items performed, either in piloting, or in live exam

Unprofessional

- Failure to accept that a “good test” can be ruined by inadequate application of suitable administrative conditions, lack of or inadequate training of markers, lack of monitoring of marking, lack of double / triple marking.

Unprofessional

- Exemption from school exams if a recognised exam has been passed. Free valid certificates should complete free valid public education.
- Use of terminology, eg “calibration”, “validity”, “reliability”, without understanding what it means, or knowing that there are technical definitions.
- Lack of acknowledgement that it is impossible to know in advance how difficult an item or a task will be.

Unprofessional

- No standard-setting: simple and naïve belief that if an item writer says an item is B1, then it is.
- No problematising of the conversion of a performance on a test of a given level to a grade result (1- 5 or A - D)

Professional?

EALTA

Present Perfect? Negative features

- Political interference
- Politicians want instant results, not aware of how complex test development is
- Politicians afraid of public opinion as drummed up by newspapers
- Poor communication with teachers and public
- Resistance from some quarters, especially university “experts”, who feel threatened by and who disdain secondary teachers

Present Perfect? Negative features

- Often exam centres are unprofessional and have no idea of basic principles and practice
- Simplistic notions of what tests can achieve and measure
- Variable quality and results
- School league tables

Present Perfect? Negative features

- Assessment not seen as a specialised field: “anybody can design a test”
- Decisions taken by people who know nothing about testing
- Lack of openness and consultation before decisions are taken
- Urge to please everybody – the political is more important than the professional

Why?

The Future

- Gradual acceptance of principles and need for application of standards
- Revision of Manual 2008
- Forthcoming Guidelines and Recommendations.
- Validation of claims: Self-regulation acceptable? Role of EALTA?
- Validation is not rubber stamping
- Claims of links will need rigorous inspection

The Future

- Change is painful: Europe still in middle of change
- Testing not just a technical matter: teachers need to understand the change and the reasons for change, they need to be involved and respected
- Dissemination, exemplification and explanation are crucial for acceptance
- PRESET and INSET teacher training in testing and assessment is essential

Good tests and assessment,
following European standards,
cost money and time

But

Bad tests and assessment,
ignoring European standards,
waste money, time and LIVES

www.ealta.eu.org

European Association for Language Testing and Assessment (EALTA)