bifie | Bildung

*standards

Testing Writing for the E8 Standards

Technical Report 2011

Otmar Gassner
Claudia Mewald
Rainer Brock
Fiona Lackenbauer
Klaus Siller

Bundesinstitut bifie

Testing Writing for the E8 Standards

Technical Report 2011

Otmar Gassner
Claudia Mewald
Rainer Brock
Fiona Lackenbauer
Klaus Siller

Bundesinstitut
bifie

Bundesinstitut für Bildungsforschung, Innovation & Entwicklung
des österreichischen Schulwesens
Alpenstraße 121 / 5020 Salzburg

www.bifie.at

Testing Writing for the E8 Standards.
Technical Report 2011.
BIFIE Salzburg (Hrsg.), Salzburg 2011

Autorinnen und Autoren:

Otmar Gassner
Claudia Mewald
Rainer Brock
Fiona Lackenbauer
Klaus Siller

Contents

Embedding the E8 Writing Test in a National and International Context

The Place of Writing in Austrian Lower Secondary Schools

There seems to be some agreement that speaking and listening are the skills most needed when trying to succeed in a foreign language environment and that being able to read is next in priority. This leaves writing as the skill least necessary for survival. Nevertheless, writing is trained from year one of secondary education on a regular basis. In some course books it starts off with model paragraphs that are personalised by the learners and leads on to open writing, mostly based on the content of the course book unit in progress. It goes without saying that lower ability learners are given more guidance, with some of them hardly ever attempting an open writing task.

In post-beginner classes the importance attributed to writing increases. It seems to be a wide-spread belief among teachers of English that when writing skills are assessed, other dimensions of language competence like vocabulary and grammar knowledge can be assessed automatically at the same time. Therefore, the writing grade goes a long way towards the overall English grade for that particular student.

Whereas this belief might be responsible for the high regard teachers have for writing, the awareness of the complexity of assessment procedures for writing is still limited. There is no perceived need for shared standard levels, there is no agreement on how writing should be tested, marked and weighted in relation to the other skills (reading, listening, speaking),[1] there are a great number of idiosyncratic marking schemes in place (even within one school), and there is no agreement on anything like pass marks or cut scores for grading.

In this situation there is room for constructive washback in the course of the introduction of E8 Standards. It is hoped that the way the tests are constructed and assessed will impact on the way writing is taught and assessed in Austrian schools.

Although much of what has been said above was formulated for the first edition of this Technical Report in 2008, it is still relevant and we can certainly see significant signs of change. A programme to train four hundred writing raters is in place and spreads expertise across the country; test specifications and a number of piloting phases have led to visible adaptations in the course books used; 'train the trainer programmes' on how to assess written performances function as starting points for school-based professional development. Finally, the reorganisation of a centralised approach to the assessment of written performances at E12 level (Matura) has contributed a lot to raising awareness of the complexity of assessing written scripts.

1 This lack of agreement is noticeable despite a clear statement in the Austrian curriculum about all four skills to be taught and trained equally in the classroom; unfortunately the curriculum does not say anything on weighting in tests. (see Lehrplan der Hauptschule. 2008, p.2)

Validity Aspects with regard to the E8 Writing Test Construct

Shaw & Weir (2007) have designed a clear graphic to illustrate their "framework for conceptualising writing test performance" (see figure 1 below). It takes all the relevant parameters into account and can serve as the blueprint for the description of the E8 Writing Tests and the theoretical framework on which they are based. Within this framework the focus of the discussion will be on the following aspects: test taker characteristics, cognitive validity, context validity, scoring validity, and consequential validity.

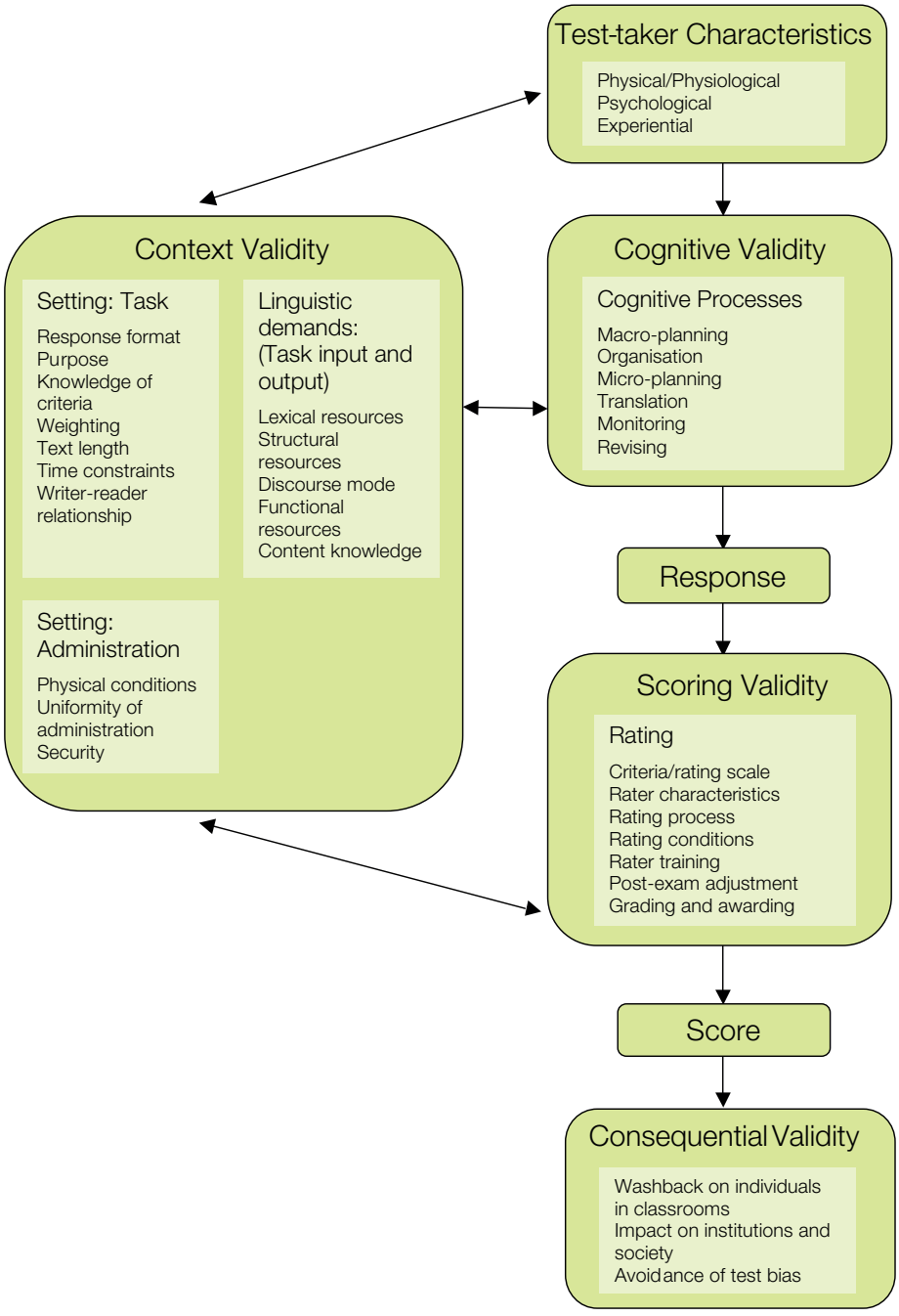

Figure 1: Adapted from Shaw & Weir 2007, 4

Test Taker Characteristics

It is obvious that test taker characteristics have an influence on the way a task is processed and a text is written. Three categories have been identified as physical/ physiological, psychological and experiential characteristics (Shaw & Weir 2007, 5).

As regards the first category, any provisions made for schooling can be considered sufficient for the E8 test situation as all test takers are pupils in the last form of lower secondary schools in Austria. To put it simply, any pupil who is fit enough to attend English classes at an Austrian secondary school and to be assessed is fit to take the E8 Writing Test.

Psychological factors, however, are almost impossible to control. Most critical is motivation as E8 Standards is a low-stakes exam that has no influence whatsoever on the individual test takers' marks or on their school career. We can expect low achievers to be more affected by lack of motivation. For this reason, test results might not fully represent the actual language competence of these students, but they might appear to be at a significantly lower level because a fair number from this group of test takers may choose not to show what they can do in English. As long as the test has no practical implications for the individual test taker, it will be difficult to generate real interest and motivation in those that display a 'could-not-care-less' attitude.

In 2013 the E8 Writing Test will be administered nation-wide for the first time. This has already had some impact on teacher attitude and might also have a positive influence on learner motivation. Preferred learning styles and personality traits are other factors that are relevant, but cannot be catered for in the given test situation.

The third group of factors are experiential characteristics referring to familiarity with the test format. Whereas the test takers are all new to this particular type of testing, they should generally be familiar with the type of prompts used in the E8 Writing Test. As details from the test specifications confirm (see pp. 20–36), prompts used are based on the *BIST-Verordnung (Anlage zur Verordnung der Bundesministerin für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen – BGBl.II Nr.1/2009 v. 2.1.2009)*, the CEFR *(Common European Framework of Reference for Languages: Learning, Teaching, Assessment)* and the Austrian curriculum (Lehrplan der Hauptschule 2008 und Lehrplan der AHS 2006).

Learners who have only done tasks that are heavily scaffolded will find the E8 prompts challenging. Those who have never faced open writing tasks in their learning history cannot be expected to perform well in the E8 Writing Tests or in international tests. We would consider it important washback if course book authors and, consequently, also teachers were to rethink the issues involved and also attempt unscaffolded writing tasks with ALL pupils. After four years of English at secondary school and some (very limited) writing at primary level amounting to more than 500 lessons, any student should be able to do a task like the one below successfully:

You have come back from a one-week stay with a host family in Cambridge. At home you remember that you left your mobile phone in your room in Cambridge. Write a short **email** to your host family.

- Tell them where you are now.
- Tell them about your mobile.
- Ask them for the mobile.
- Tell them how you liked your stay.

Figure 2: BIFIE Item Archive (http://www.bifie.at/freigegebene-items)

2011 is the first year with a new generation of course books available for Austrian schools to choose from. What was formulated above as expected washback in 2008 has materialised: The new course books include writing tasks that are geared to the E8 Writing Specifications with a number of them extremely close to actual E8 Writing Prompts. Even the time constraints and the specifications regarding length have been taken on board. Another salient feature is the attempt to actually teach the students about using paragraphs when producing (longer) texts.

Cognitive Validity

"The cognitive validity of a Writing task is a measure of how closely it represents the cognitive processing involved in writing contexts beyond the test itself, i.e. in performing the task in real life" (Shaw & Weir 2007, 34). Whereas it is notoriously difficult to describe the cognitive processes as they are not directly accessible, it seems important to describe a general writing model that accounts for writing in a real-life context as well as in an exam situation. However, one difference should be noted at the outset, namely that there is no time-constraint in most real-life situations whereas in the E8 testing situation time, topic, genre, and length of output are predetermined. This might impose limitations on the planning phase as well as on the writing and revision phases.

Writing Theory in Brief

In the given context, only sketchy references shall be made to various sources that present and discuss the writing process and models of L1 and L2 writing in detail. According to Grabe and Kaplan (1996, 230–232), the planning phase, which they call "goal setting", involves these five factors:

- an assessment of the context
- a preliminary representation of the writing product
- an evaluation of possible problems in task execution
- an initial consideration of the genre required
- an organisational plan

Shaw and Weir (2007, 37) make a point of emphasising the advantages of a more psycholinguistically oriented model of writing over the Grabe and Kaplan model and refer to Field (2004) and Kellogg (1994, 1996). Interested readers may wish to consult the detailed discussion there. The Field model (Field 2004, 329–331) involves

- macro-planning
- organisation
- micro-planning

- translation
- execution
- monitoring
- editing and revising

A reference to Scardamalia and Bereiter (1987) is essential here as they have described two different strategies used by skilled and less skilled writers in the planning phase: *knowledge telling* and *knowledge transformation.*

> In knowledge telling, novice writers plan very little, and focus on generating content from within remembered linguistic resources in line with the task, topic, or genre. Knowledge transforming by the skilled writer entails a heightened awareness of problems as and when they arise – whether in the areas of ideas, planning and organisation (content), or in those of goals and readership (rhetoric) […] (Shaw & Weir 2007, 43).

Whereas this holds true for all writing, L2 writing poses additional cognitive demands on the writers as Field (2005) argues. Attention directed towards linguistic aspects like lexical retrieval, spelling, and sentence structures can impede the fluency of writing and the capacity to organise and structure the text as a whole. Some ideas might have to be abandoned in the execution phase on the grounds of language constraints and limitations.

Cognitive Processing in the E8 Writing Test

In the E8 context we suggest using a modified Grabe/Kaplan-Field model to illustrate the writing process, which will clearly be based on *knowledge telling* and thus has a very brief planning phase mainly consisting of considering relevant content points.

This model includes the following phases:

- assessment of the context (who writes about what to whom and why?)
- characteristic features of the genre required
- preliminary representation of the writing product
- selection of content points
- evaluation of possible problems in task execution
- micro-planning at paragraph and sentence level
- translation
- monitoring
- revising

Figure 3 is a graphic representation of the modified Grabe/Kaplan-Field model highlighting the three main steps.

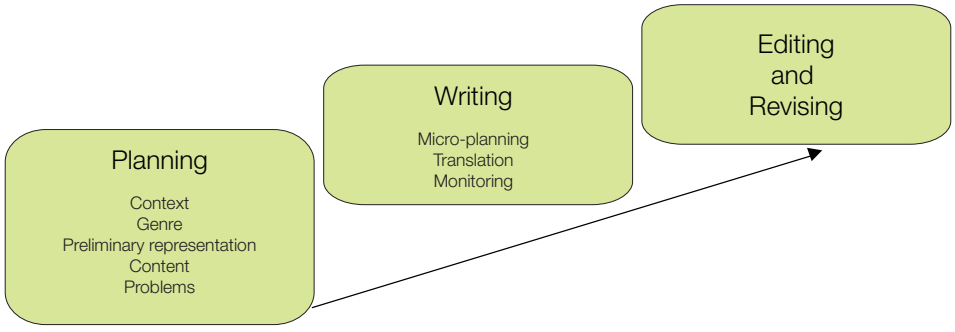

Figure 3: Modified Grabe/Kaplan-Field Model

In the E8 test situation, the planning phase is clearly not elaborate or extensive. After an assessment of the context, which includes identifying the topic, the situation of the writer, the text purpose and the addressee, most test takers will move straight to the consideration of the genre required and develop a "preliminary representation of the writing product". Then the bullet points will pre-structure the content elements to be included. An organisational plan is not necessary as the tasks are fairly short and guided by content points with little opportunity for deviation. Especially with the short task (40–70 words) planning efforts will be reduced to the bare minimum and be restricted to the decision on which content points to elaborate and how to proceed in that direction.

The writing is more closely guided than in real life as a number of content points are given in the prompt. This makes the writing process somewhat easier than in real life, but on the other hand, it seems unavoidable if we want to ensure inter-rater reliability for the dimension of task achievement. In addition to this, providing a content schema for candidates at this level is necessary because the cognitive load for simultaneous activities on a macro and micro level would be too great and the task too demanding.

It has become clear from the present discussion that macro-planning and organisation play no role in the given writing context and that the product delivered will be firmly set in the area of *knowledge telling*.

The micro-planning phase, the next step of the writing process, might be the point where possible problems in task execution will be identified before the actual writing begins. The problems will be content-related and have to do with knowledge of the world and what (abstract) ideas to use with the content points given; they might also be connected to the attempt to recall the requirements of the genre in question and with the language necessary to express the ideas.

This stage of identifying language resources and their limitations is only a fraction away from actually putting pen to paper and undoubtedly is a central aspect of micro-planning focusing

> on the part of the text that is about to be produced. Here, the planning takes place on at least two levels: the goal of the paragraph, itself aligned with the overall goal of the writing activity; within the paragraph, the immediate need to structure an upcoming sentence in terms of information (Shaw & Weir 2007, 39).

Micro-planning merges with the translation phase where previously abstract ideas only accessible to the writer him/herself are translated into the public space defined by a shared language. In contrast to Shaw & Weir and Field, we see micro-planning and translation as two stages that are interlinked as the writer might oscillate between the one and the other at sentence level or at paragraph level (Shaw & Weir 2007, 39–40).

It is in the translation stage that language decisions have to be made and planning decisions have to be implemented. The actual production of text takes place under the constraints of content schemata, genre restrictions and the limitations of linguistic resources at hand in L2. What has been called "avoidance behaviour" (e.g. avoiding lexis or structures that seem unsafe) and "achievement behaviour" (e.g. using simpler structures, paraphrasing) by Field (2004, 66–67) needs to be taken care of in the assessment phase, as does the ability to produce coherent and cohesive texts.

The next step is monitoring although this is not necessarily sequential and might be oscillating with phases of translation. "At a basic level monitoring involves checking

the mechanical accuracy of spelling, punctuation and syntax" (Shaw & Weir 2007, 41). At E8 level this is what can be expected, if not in the lowest segment of test takers. In addition, better writers will also check back on content and genre requirements. These monitoring activities will lead to editing and revising if some parts of the text have been found unsatisfactory. This might involve adding, deleting or modifying a content point, adding cohesive devices, replacing poor words and phrases with better ones, or simply correcting mistakes in spelling and structure.

In the E8 context, writing is certainly based on the *knowledge-telling model* (Scardamalia & Bereiter 1987); Hyland's summary of the model epitomises E8 writing performances:

> **A knowledge-telling model** addresses the fact that novice writers plan less often than experts, revise less often and less extensively, and are primarily concerned with generating content from their internal resources. Their main goal is simply to tell what they can remember based on the assignment, the topic, or the genre (Hyland 2002, 28).

Context Validity

Tests should be as close as possible to authentic real-life situations. Writing is an activity that is normally performed by individuals at a time set aside for it. Writers have a purpose and an audience; they have the freedom to interrupt the writing process and resume it at a time of their choice, especially for editing and revising; and they can normally use dictionaries and other resources. In the given test setting, some constraints will be operative, but unavoidable.

Shaw & Weir 2007 (64–142) discuss a number of aspects of context validity related to three areas:

- Setting: Task
- Setting: Administration
- Linguistic Demands: Task Input and Output

These points will structure the discussion of context validity of the E8 Writing Tests.

Setting: Task

The aspects to be discussed here are response format, purpose, knowledge of criteria, weighting, text length, time constraints, and writer-reader-relationship. In the E8 Writing Tests authenticity is one of the most prominent aims of prompt construction. However, in contrast to real-life writing there is no provision for the use of any resource materials such as dictionaries.

The writing tasks are targeted at pupils of Austrian schools in year 8 and normally aged fourteen. The tasks are designed to appeal to this age group and to elicit scripts that show what test takers can do within the framework defined in the *BIST-Verordnung*. The domains and genres have been carefully selected from this framework, which is based on the CEFR, and have been filtered further on the basis of the Austrian curriculum.

As the response format may well play a significant role in test performance (Alderson et al. 1995), the decision has been taken to include two formats in the E8 Writing Test. There is a short task (40–70 words) and a long task (120–180 words), which are assessed separately. Both are open writing tasks. Good writers have a better chance to show their best in the long task, which is based on a B1 descriptor, taken from the

BIST-Verordnung. Lower achievers are expected to do better in the short task, which is limited in scope, more closely guided and based on an A2 descriptor. However, both good and weaker writers are expected to address both tasks as they are not supposed to choose only one of the tasks.

Instructions, delivered both orally and in writing to the test takers before the actual test by a test administrator, and rubrics that go with each task present candidates with information regarding text length (see above) and time constraints. For completing both tasks the test takers have 30 minutes of writing time plus 5 minutes for editing and revising in all. After 35 minutes there is some time for word count by the candidates. The actual prompts contextualise the task by defining the writer-reader-relationship, stating purpose and genre, and giving content points to be included in the text. The short task contains 3–4 content points, the long one 5–8.

Information on the scoring criteria used and their weighting, including the rating scale used, scale interpretations and benchmarked sample scripts, is published in this report (see pp. 12–14, 23–27, 28–36, 37–44). Furthermore, sample prompts, the rating-scale and benchmarked texts are publicly available on the BIFIE website[2].

Setting: Administration of E8 Writing Test

In its present form, the writing test was first piloted on a sample of ca. 800 test takers in 2007 and in 2009 a baseline study was carried out. Consequently detailed information on the "pilot phase" between 2006 and 2008 and on the baseline tests in 2009 were published in a Technical Report (Breit & Schreiner 2010). Starting in 2013, the E8 Writing Tests will be set nationwide every three years and all Austrian school children in grade 8 will be tested. Only SEN pupils, i.e. those with special educational needs, will be exempted from doing the tests.

In order to ensure reliable test results, the circumstances under which the E8 Writing Test takes place must be similar. The steps discussed in more detail here concern physical conditions, uniformity of administration, and test security, based on the ideas by Shaw & Weir (2007).

As the venues of the E8 Writing Test are classrooms in Austrian schools, physical test conditions are of very similar standards and test takers should find appropriate conditions for taking the test.

In order to grant the uniformity of administration, the test must be conducted according to standardised instructions by trained test administrators. An extensive test administrator's manual is provided during the test administrator training. The manual includes information on the background of the test, checklists and To Do's both for the preparation, the actual setting of the test (e.g. starting the exam, completing different lists, standardised verbal instructions for the test administrator etc.), and the conclusion of the examination.

In a nationwide exam there are some administrative constraints: a political decision has been taken regulating test administration in the years to come: in 90 % of the classes the E8 Tests will be administered by the teachers of the school (internal test administration). In a further 3 % of the classes the tests will also be administered internally, but there will be external quality monitors to assure the correct and standardised administration of the tests. 7 % of the classes will be tested externally. All test administrators, both internal and external ones, are trained to administer the

2 http://www.bifie.at/freigegebene-items [24 June, 2011]

tests according to agreed standardised procedures. However, it is within the responsibility of the schools' head teachers to take care of a correct and standardised test administration, as this is the only way to get reliable feedback regarding the performance of their pupils and to plan local measures of quality development.

The prompts used in future writing tests have all been written by the prospective raters, moderated, edited, and screened by the BIFIE Writing Trainer Team, pretested and stored in the item archive. The test booklets are designed by the same group in cooperation with the psychometric department at BIFIE. The actual distribution of all test papers to the schools is handled centrally by BIFIE Salzburg.

More detailed information on the administration of the E8 Tests will be published in a Technical Report after the first nationwide testing in 2013.

Linguistic Demands: Task Input and Output

In the Austrian teacher community the communicative approach to language learning (Canale & Swain 1980 as an important precursor of the Bachman 1990 model of communicative language ability) is widely accepted, and it is also set down in writing in the national curriculum. As the learning tasks are modelled on real-life contexts, the learning environment aims to mirror real life as closely as possible. Exams set in the Austrian context need to share these premises and to reflect them in the tasks set.

Shaw & Weir (2007, 91), Alderson (2004, 13) and others complain that the CEFR remains vague and withholds details when it comes to structures or vocabulary, using terms like "simple" in the descriptors. While this is true, reading the CEFR extensively rather than focusing only on the sections containing the scales proves helpful. In chapter 3, the development of the common reference levels is explained and it is made clear that they progress in a very coherent way from "the lowest level of generative language use" (CEFR 2001, 33) to social functions and "descriptors on getting out and about" (CEFR 2001, 34) based on *Waystage* (A2) and a simplified version of some transactional language from "'The Threshold Level' for adults living abroad" (CEFR 2001, 34). A2+ does not so much increase the range of topics, but focuses on "more active participation in conversation" and "significantly more [on the] ability to sustain monologues". B1 reflects the *Threshold Level* and involves "the ability to maintain interaction and get across what you want to, in a range of contexts" as well as "the ability to cope flexibly with problems in everyday life" (CEFR 2001, 34). B1+ makes increased demands on the quantities of information to be handled.

As this is the way the levels have been constructed (i.e. from *Waystage* to A2), it seems legitimate to move from A2 specifications back to *Waystage*. And here we have a vocabulary list and a list of structures considered characteristic of that level. As UCLES have also oriented themselves on vocabulary lists from the Council of Europe Publications (Lexical Inventory in *Waystage,* 1980, 45–62; and in *Threshold,* 1979, 85–115), it can be considered a useful shortcut to pick up the vocabulary lists published on the web for KET (A2) and PET (B1), especially as these have been updated on the basis of frequency and usage data from language corpora. Generally, "[the] language specifications of KET are the same as those set out in Waystage 1990" (KET Handbook 2007, 1).

To resume the discussion of the vagueness of descriptors using words like "simple", "basic" or "sufficient", it may suffice to say that this vagueness needs to be contextualised. If the A2 descriptor on Grammatical Accuracy reads "Uses some simple structures correctly, but still systematically makes basic mistakes" (CEFR 2001, 112), we can expect learners to use the range of structures listed in the *Structural Inventory of*

Waystage (63–83) or the *KET Handbook* (8–9) with severely restricted accuracy. In this sense, even vague terms like "simple" are reasonably well-defined so that prompt writers and raters know what to look for.

The E8 writing prompts do not restrict test takers in their use of specific lexical or structural resources, but give them the opportunity to demonstrate their linguistic abilities within the task set. The extent of their success in doing so is assessed according to the graded descriptors in the assessment scale.

What has to be noted, however, is the basic orientation of the CEFR towards an adult learner and a dominance of tourist aspects of language learning. This is why the Austrian E8 Standards have also integrated the specifications set down in the Austrian curriculum and adapted the CEFR descriptors to the age group of the test population. This mainly reflects the selection of domains and transactional situations. It has no influence on the structures included, though it has some influence on the wordlist. Generally, the school books used in Austria take this into account. As the test is explicitly based on the Austrian curriculum, the linguistic demands of the test are fair for all test takers.

The writing prompts used often specify particular language functions to be performed, e.g. "invite…, apologise…, ask for…, give advice…". A list of these functions has been made available to the teachers preparing the test takers so that they can be expected to be aware of them (see pp. 53–54 in the appendix).

Several research papers have observed an interaction or even an interdependence of content knowledge on the one hand, and writing performance and test scores on the other (Read 1990, Papajohn 1999, Weir 2005). Provisions for this have been made by restricting topics to areas that can safely be assumed to be familiar to the test takers as they are set down in the Austrian curriculum and must have been included in their English lessons. However, this still leaves the fact that some test takers might feel indisposed to deal with a particular topic for a number of reasons, the most common probably being lack of motivation and interest.

More detailed information on discourse mode (i.e. text types), functional resources (i.e. intention/purpose), and content knowledge (i.e. topic area) can be found in the tables on pp. 21–22 representing the E8 Construct Space.

Scoring Validity

> Scoring validity is concerned with all the aspects of the testing process that can impact on the reliability of test scores. [… It] is criterial because if we cannot depend on the rating of exam scripts it matters little that the tasks we develop are potentially valid in terms of both cognitive and contextual parameters. Faulty criteria or scales, unsuitable raters or procedures, lack of training and standardisation, poor or variable conditions for rating, inadequate provision for post exam statistical adjustment, and unsystematic or ill-conceived procedures for grading and awarding can all lead to a reduction in scoring validity and to the risk of construct irrelevant variance (Shaw & Weir 2007, 143–144).

In this section we examine each of the relevant parameters in some detail: criteria and rating scale, rater characteristics, rater training, rating process, rating conditions, post exam adjustments, and grading.

Criteria and Rating Scale

Before the actual construction of the rating scale, information on existing scales was collected and the usefulness of the scales in the framework of E8 Testing was analysed: Jacobs et al. scoring profile 1981 (Weigle 2002, 116); TEEP attribute writing scales, Weir 1990 (Weigle 2002, 117); FCE Scoring rubric 1997 (Weigle 2002, 152); TOEFL writing scoring guide 2000 (Tankó 2005, 125); IELTS bands 2002 (Weigle 2002, 159); Analytic writing scale developed by the Hungarian School-Leaving English Examination Reform Project 2005 (Tankó 2005, 127).

Lumley reports findings from Weigle 1994, who used an analytic scale to have 30 compositions assessed by novice and expert raters. Weigle focused on novice raters, which is relevant to the E8 situation in Austria where a rating culture is only just evolving.

> She found that rater reliability increased as a result of training, and that the improved agreement was the result of raters gaining better consensual understanding of the terms and levels represented in the scale. She found evidence that training helped clarification of the rating criteria (Lumley 2005, 44).

This supports the view of the testing team that in the given context an analytic scale would be preferable to a holistic scale. This view is also supported by Weigle 2002, who mentions several advantages of analytic over holistic scoring:

- It is more useful in rater training as inexperienced raters can more easily understand and apply the criteria in separate scales.
- It is particularly useful for second-language learners, who are more likely to show a marked or uneven profile.
- A scoring scheme in which multiple scores are given to each script tends to improve reliability (Weigle 2002, 120).

Another reason for ruling out a holistic approach was the fact that rating procedures for scripts within the Austrian school system are not regulated, show great variety and are to a large extent holistic, even impressionistic. As assessment procedures for writing in Austrian schools cannot be taken as a basis for a disciplined approach towards rating scripts, breaking with this tradition seemed to best guarantee a fresh approach to assessment.

Taking the general background of Austrian traditions in assessing writing into account and inspired by the Hungarian scale (Tankó 2005, 127), the decision was taken to design an analytic scale measuring four dimensions: Task Achievement, Coherence and Cohesion, Grammar, and Vocabulary. Whereas three of these four dimensions have a strong recognition value for Austrian teachers, Coherence and Cohesion might appear unusual and reflects the high importance given to this dimension by the CEFR. These four dimensions promised to yield enough detail for a constructive feedback profile on individual test taker performance, information for instruction as well as informative data for system monitoring.

The assessment scale was constructed bearing in mind the fact that the overall majority of performances could be expected to be around A2/B1. This meant that A2 and B1 descriptors needed to be included while anything at and above B2 could be neglected. We are aware of the fact that this kind of scale cannot measure B2 or C1 performances and we have settled for stating that performances above the upper end of the descriptors in the E8 Scale are called “above A2” for short tasks and “above B1” for long tasks. But, generally, the applicability of a particular descriptor does not

automatically signal that a script is at a given CEFR level. Firstly, bands consist of more than one descriptor, and secondly, linking written performances to the CEFR is a complex procedure that is beyond the scope of this report and will be discussed in a separate publication.

The second consideration in scale construction was the cognitive load that raters can manage in the rating process. The decision to use four dimensions is also in agreement with the CEFR recommendation to reduce the number of possible categories to "a feasible number" as "more than 4 or 5 categories starts to cause cognitive overload" (CEFR 2001, 193). We take it that this warning also applies to the number of bands and descriptors that raters can handle, so we have opted for four bands supplied with descriptors and three empty bands in between, making it a seven-band scale plus a zero band.

At that point in scale construction the scales consisted of three columns: The first being a deflated descriptor for each of the four bands, the second being extended and containing more detail, and the third quoting the related CEFR descriptor. An important decision in the process of scale construction was the removal of the CEFR levels at the end of the CEFR descriptors and, in a second step, the removal of the CEFR descriptors altogether. This was the logical step to take when some raters awarded band 7 to a script and argued that the script was a B2 performance. However, such an argument is inadmissible as the prompts used in the test are written on the basis of A2 or B1 descriptors and responses to these prompts simply cannot represent performances above A2 or B1 respectively as one basic factor is the scope of a performance together with the given limitations of domains and genres. So even when the CEFR B2 descriptor for Grammatical Accuracy "Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding." (CEFR 2001, 114) describes the performance well, it does not mean that it is B2, but that the A2/B1 task has been carried out very well and that the (grammar) performance is a very good A2 or B1 performance respectively.

In another step, the scale was condensed to one page with an extended scale each for Task Achievement Short and Task Achievement Long. As this deflated scale might not carry enough information for the raters at the beginning of their training, scale interpretations have been provided (see pp. 37–44). The scales themselves have been fine-tuned in the training process in an ongoing dialogue with the raters. It follows from this that the scales are what has been called assessor-oriented (Weigle 2002, 122; CEFR 2001, 38).

The writing scripts are assessed on four dimensions: Task Achievement, Coherence and Cohesion, Grammar, and Vocabulary. Whereas the last three are based on the CEFR and the Austrian *BIST-Verordnung,* the CEFR does not contain anything explicit on Task Achievement. The descriptors of the scales on *Overall Written Production* and *Overall Written Interaction* mainly refer to linguistic and pragmatic aspects *(Can write a series of simple phrases and sentences linked with simple connectors …),* whereas the subscales only make references to text types, domains and content aspects *(Can write personal letters describing experiences …).* These descriptors cannot be operationalised in assessment terms.

In our view, however, the content aspect of writing is central and largely responsible for the overall quality of a script. Nevertheless, the raters do not give an overall grade for writing, but all four dimensions are rated separately and are reported as a profile, which more often than not is uneven or marked. For the feedback procedure, an overall writing score with the four dimensions of the short and the long performance assessment is given, based on an equal weighting of all dimensions.

Rater Characteristics

It has been reported that "Subject specialists and language trained EFL teachers demonstrate a tendency to employ rating instruments differently" (Elder 1992, in Shaw & Weir 2007, 169). In this respect the present situation in Austria is uncomplicated as all raters are teachers of English who teach in lower secondary schools. Some of these are native speakers now living and working in Austria, some have a university background, others were educated at University Colleges of Teacher Education.

Although the raters go through a specific training that familiarises them with the rating scales and the rating procedures, differences in their experiential background and in their professional training and development may lead to differing assessments of scripts. In order to make raters aware of this and to start a process of self-reflection, all raters get detailed feedback on their rating behaviour at several points in the training and particularly after the last training session and after the administration of a writing test. They are informed about their inter-rater reliability and rater severity. Eventually, harshness and leniency of raters is taken care of through Rasch modelling.

Rating Process

Milanovic et al. (1996) identified a number of approaches raters take in the process of rating a script. In our training sessions we generally advise against the "read through" and the "provisional mark approach", both of which are based on one reading of the script. Raters are encouraged to adopt a "principled two scan/read approach" to the process with a focus on Task Achievement and Coherence and Cohesion in the first reading and on Grammar and Vocabulary in the second. The length of the scripts seems to support this approach.

We are aware of group effects on rater reliability as described by Shaw & Weir (2007, 174–175) and have made an effort to use them to our advantage in the standardisation meetings at the beginning of the training sessions and the rating session. In addition to the procedures recommended for standardisation meetings (Alderson, Clapham & Wall 1995, 112–113) a considerable amount of time is spent on the detailed interpretation of the prompts (see appendix, pp. 48–52) and an open discussion of any questions that might be raised by the raters taking into consideration that all raters have also been involved in the writing of prompts and their piloting. An additional set of ten benchmarks, gained in an extensive benchmarking conference with ten benchmarkers, plays a vital role in the standardisation sessions.

Rating Conditions

In 2013 the whole E8 population of some 90,000 pupils will be tested. In June, all raters, who have been trained at different intervals since 2006, will take part in a one-day standardisation meeting as described on p. 18 in training phase 6, followed up with a one-day rating session.

There will be regional standardisation meetings for all raters who mark scripts from the E8 Writing Tests. In these sessions raters are updated on, for example, any changes regarding the assessment scale used. Then they will continue with the on-site rating session, in which they will undergo supervised rating with the new test prompts and actual scripts of the test. The scripts have been carefully compiled in rating booklets by the BIFIE Salzburg psychometric department. This will provide BIFIE with the relevant data needed for test analysis and feedback.

It will take half a day's work to deal with each prompt and give raters enough time for on-site rating and clarification of rating problems based on the particular test prompts. The remaining unrated scripts, approximately three quarters in total, will be rated off-site within six to eight weeks at the raters' convenience.

Rater Training

According to Alderson, Clapham & Wall, rater training is one of the most essential aspects in an effort to obtain reliable assessments (1995, 105). Lumley refers to „a growing body of work that shows the effectiveness of the training process in allowing raters to develop adequate reliability or agreement in the use of individual scales in relation to specified test tasks" (2005, 62).

This has been taken very seriously by the BIFIE Writing Trainer Team, who have developed a seven-months training programme for raters starting in October and preparing the raters for the mock rating session in April/May. This programme is described in some detail below.

RECRUITMENT
In the recruitment phase teachers in Austrian lower secondary schools are approached to become writing raters. As the test takers come from the two different types of lower secondary schools, the General Secondary School (Hauptschule and Neue Mittelschule) and the Academic Secondary School (AHS), care has been taken to ensure intake of raters from all three of these school types. While recruitment was originally carried out by BIFIE Salzburg until 2009, the administration of the recruitment process has since been outsourced to the regional University Colleges of Teacher Education.

TRAINING PHASE 1: OCTOBER (1 DAY; FACE-TO-FACE SESSION)
As the CEFR is the most relevant background document for the E8 Standards, the starting point of the first training session is *The Common European Framework* in general and the *Overall Writing Scales for Production and Interaction* in particular. The familiarisation with the CEFR is implemented on the basis of the recommendations made in the *Manual on Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF) (2003, 70–77),* including sorting tasks. It is made clear at this point that one aspect of writing is related to communicative activities and strategies, another one to linguistic competences.

The Test Specifications are presented and discussed in detail: domains, genres, construct space, prompt format. In this phase there is a focus on prompt production. There are *Guidelines for prompt writers* that provide assistance in the process of prompt writing.

Each prospective rater takes on the task of writing one long or one short prompt in a training tandem during the weeks following the first training session. All prompts are sent in to the trainer team for moderation. Once the prompts have been screened, they are piloted by the prompt writers. Behind this procedure there is the belief that raters need to know about the qualities of prompts and what elements they must contain. This prepares them for better interaction with the test prompts in the actual rating phase.

With regard to differential validity it is important to state that every precaution is taken at the early stage of prompt writing to avoid test bias. Varying cultural backgrounds and knowledge of the world have been taken into account as well as the given variety of cognitive characteristics, mother tongue, ethnicity, and gender.

After an introduction to the Austrian E8 Standards the rating scale is presented and explained on the basis of written scale interpretations (see pp. 37–44). Unfolding the seven bands with four of them defined and working through the four dimensions takes time. The procedure adopted is to look at the seven bands of one dimension, which leads to some theoretical understanding of the scales, but connections to actual scripts are still tenuous. So the raters get two scripts each for individual marking on the first dimension (task achievement). They discuss their assessment in small groups and the trainer discloses the benchmarked assessments and argues the case. This procedure is repeated for the other three dimensions.

In the second phase of the training workshop the participants get sets of benchmarked scripts which they rate on all four dimensions. Benchmark justifications are presented in detail and discussed in plenary to ensure that the raters understand the descriptors of the rating scale and can handle the rating agreements that have been formulated for a number of special rating problems (e.g. downgrading for below-length texts). After discussions and argumentation of the judgements for the benchmarked scripts the participants have a rough idea of the complexity of the rating process and the effort it takes to arrive at reliable judgements.
The rating sheets filled in by the participants provide a first set of data that helps to monitor intra-rater reliability.

TRAINING PHASE 2: OCTOBER – DECEMBER
The second training phase is an open one with a fairly loose structure. All participants first write their prompts and get them back from the testing team as screened prompts (in some cases the prompts are returned to the writers for repair). As a first measure towards quality assurance the prompt writers produce a response to their own prompt. This should make prompt writers aware of the more obvious flaws their prompts might have. The prompts are then piloted in one of their classes so that all participants have around 20 scripts based on their prompt.

TRAINING PHASE 3: JANUARY – MARCH
Once the prompt writing and piloting is finished, the online rating phase starts in January with the trainer team sending out scripts to the raters for individual rating on each of the four dimensions. The raters have about four weeks to do this and send in their ratings. When all scoring sheets have been submitted to the testing team, the benchmarks are sent out to the raters. In February/March the raters practise their rating skills on their own pilot scripts. They select two scripts to be fed into the training process. They rate these scripts and write justifications for their ratings. A second online rating phase helps to standardise the raters, who are encouraged to make final adjustments to their scores and judgements before sending them to the trainer together with the digitalised scripts. The trainer goes through these scripts and selects interesting samples for the upcoming training workshop in April/May.

TRAINING PHASE 4: APRIL/MAY (1 ½ DAY; FACE-TO-FACE SESSION)
Training Phase 4 starts with a discussion of open questions from previous training phases. Then there are two standardisation sessions with recycled scripts and new benchmarks and the time is spent rating scripts and discussing particular problems arising in the process.

After these two standardisation sessions the first prompt writer tandem presents their prompt and the whole group rates 2–4 scripts. Experience from previous rating sessions has shown that, as the raters have to handle a number of different prompts in this phase, they need more guidance in the analysis of the prompts and therefore are also provided with prompt interpretations. The prompt writers then disclose their judgements and defend their scores in a discussion with the whole group monitored

by the trainer. This procedure is repeated so that the majority of the raters have the chance to discuss their scripts and their judgements with the whole group. Inter-rater reliability and intra-rater reliability are monitored and pertinent data on each rater is collected systematically.

TRAINING PHASE 5: MOCK RATING SESSION: IMMEDIATELY AFTER TRAINING PHASE 4 (1 ½ DAY; FACE-TO-FACE SESSION)
The second important part of this meeting is the analysis and interpretation of new prompts that were piloted on a representative sample of prospective Austrian test takers and which might actually be used in a future test. Raters are given detailed information and are invited to discuss any issues that are still unclear.

Then the actual rating begins. The raters receive booklets of piloted scripts which were written in response to a short or a long prompt. There is a rating plan with overlap for multiple rating. After about ten scripts have been rated, the raters meet with the supervisor to discuss any critical issues that may have come up during the rating. Then they proceed to rate the other scripts of that booklet, which involves some free time management for the raters. This procedure is then repeated for the other booklets.

The scoring sheets filled in by the participants provide a set of data for the analysis of rating behaviour. The data are used to give extensive feedback to all raters on their inter-rater reliability and rater severity.

TRAINING PHASE 6: UPDATE SESSION IN THE YEAR OF THE ACTUAL TEST (1 DAY; FACE-TO-FACE SESSION)
There are regional standardisation meetings for all raters who mark scripts from the E8 Writing Tests. In these sessions raters are updated on, for example, any changes regarding the assessment scale used. Then time is spent on the analysis and interpretation of the prompts used in the actual test, and benchmarked scripts based on these prompts are rated.

After the update and the standardisation session, the on-site rating session starts, following the same procedure as described in training phase 5 above.

Post Exam Adjustments

Although considerable efforts are taken in the training programme to minimise discrepancies in rater behaviour, the ratings are adjusted for any remaining differences in rater severity by means of multifaceted Rasch analysis after the scripts have been marked. This becomes possible by having a certain proportion of scripts marked by two raters (double rating) and another proportion of the scripts by all raters (multiple rating) so that rater behaviour can be assessed in terms of model fit as well as severity.

Reporting Results

The purpose of the E8 Standards is giving feedback on the writing competence of Austrian pupils in grade 8. The aim, therefore, is system monitoring rather than certification or selection at the level of individual test takers. Cut scores will be established, however, to enable individual feedback to test takers and show whether the objectives in the national curriculum have been met as regards writing competence in general and all four dimensions in particular. Consequently, while the test results are linked to the CEFR, critical cut scores on which to base selection decisions need not be established by the test constructors. It is hoped that by providing results to individual teachers and schools this feedback will instigate a qualitative development

that will radiate beyond regions and spread throughout the whole school system. The way feedback on the results is given to test takers and other stakeholders is being developed at the moment. In compliance with political requirements, only the test takers themselves will have access to their individual results through a code they will be given when sitting the exam. Teachers and school principals will receive aggregated data for the group relevant to them (class, school) via an internet platform. Educational authorities will receive aggregated reports.

The information that results from the writing test is reported on the four dimensions of the Writing Scale (Task Achievement, Coherence and Cohesion, Grammar, Vocabulary). The results for each dimension are reported on a scale from 0 to 7, which enables reference to the CEFR up to B1. Ratings are adjusted for differences in rater severity and task difficulty by means of multifaceted Rasch analysis. The results are therefore comparable across all test takers regardless of which rater rated the performance and what particular prompt the performance is based on. The process of standard setting and CEFR-linking will be described in more detail in a technical report after the actual test in 2013.

Consequential Validity

Shaw & Weir (2007, 218) take the term 'consequential validity' from Messik 1989 and interpret it in the light of recent literature to include washback (influences on teaching, teachers, learning, curriculum and materials) and impact (influences on the community at large). The E8 Standards can be envisaged as an instrument to initiate changes in the direction of positive or beneficial washback.

In and around 2008 new course books for teaching English to the target group were launched and a number of them claim to be informed by the CEFR and the E8 Standards. This means that text book writers are well aware of the E8 Standards Tests and are adapting their materials towards them.

The requirements for the writing test are clearly laid down in this report and demonstrate what kinds of writing our learners are expected to deliver.

The expectations of the test designers formulated in 2008 have been largely fulfilled[3]. Three years later a great number of writing tasks in the new course books used in Austrian schools have changed in the direction indicated in the Technical Report 4 of 2008. There is much less scaffolded writing; the tasks are realistic and authentic; text type requirements, variation in text types, text length and time constraints are all in line with the present test specifications. Some course books also emphasise the use of paragraphs in writing, give hints on how to write good paragraphs, and provide corresponding exercises. This means that Austrian test takers who sit the E8 Tests after 2011 will be familiar with the test format, the particular requirements, and the instructions.

3 First edition, 2008, p.19: „It is hoped that this will lead to less scaffolded writing, thus enhancing learner empowerment. The emphasis given to coherence and cohesion in the CEFR and the E8 Standards might also focus teacher attention on this area and entail improvements."

E8 Writing Test Specifications Version 03 (July 2011)

The guiding documents for the development of the writing test specifications for the E8 Writing Test are the Austrian curriculum (AHS 2006; APS 2008), the *BIST-Verordnung (BGBl.II Nr.1/2009 v. 2.1.2009)* and the CEFR (Council of Europe, 2001). The first two documents list writing competences at different proficiency levels in terms of the CEFR.

1. Purpose of the Test

The main purpose of the writing test is to identify strengths and weaknesses in test takers writing competence and to use this information both for the improvement of classroom procedures and for system monitoring. What is more, individual and detailed test results are reported to the test takers, which is of interest to the test takers themselves and their parents.

2. Description of Test Takers

The test takers are Austrian pupils in the two different types of lower secondary schools, the General Secondary School (Hauptschule and Neue Mittelschule), and the Academic Secondary School (AHS), towards the end of grade 8 (8. Schulstufe). Pupils from all three ability groups in APS will be tested. The majority of test takers will be aged 14. SEN pupils, i.e. those with special educational needs, will be exempted from doing the tests.

3. Test Level

The difficulty level of the test is supposed to encompass levels A2 to B1 in the CEFR.

4. Test Construct with E8 Construct Space

The tables on pp. 21–22 summarise the construct space relevant for item design with a list of the prompt types used to test the writing competences as specified in the *BIST-Verordnung,* targeted at levels A2, A2+, and B1 of the CEFR. The tasks at these levels ask for (mostly) concrete content. Therefore topics are restricted to areas that can safely be assumed to be familiar to the test takers as they are set down in the Austrian curriculum and must have been included in their English lessons.

More specifically, the tasks display various text types and writing intentions/purposes. Tables 1 and 2 provide an overview of the range of text types and writing intentions/purposes for the proficiency levels tested. For the actual construction of writing items prompt writers are given special prompt design specifications, which clearly list what kind of prompt – in terms of prompt type, level, BIST-Descriptor, topic area, and text type – the prompt writers are supposed to create.

E8 Construct Space

Prompt Type	CEFR Level	CEFR Descriptor	Deskriptor aus BIST-VO: Schüler/innen können…	Topic Area	Text Types	Intention/ Purpose	Primary Audience
Long Prompt	B1	▪ Can write accounts of experiences, describing feelings and reactions in simple connected text. ▪ Can write a description of an event, a recent trip – real or imagined. ▪ Can narrate a story. ▪ Can write personal letters describing experiences, feelings and events in some detail.	▪ Erfahrungsberichte schreiben, in denen Gefühle und Reaktionen in einem einfachen, zusammenhängenden Text wiedergegeben werden ▪ eine Beschreibung eines realen oder fiktiven Ereignisses, z. B. einer Reise, verfassen ▪ eine Geschichte erzählen ▪ ausführlichere Karten, persönliche Briefe und E-Mails schreiben und darin auch über Ereignisse, Erfahrungen und Gefühle berichten	▪ Familie und Freunde ▪ Wohnen und Umgebung ▪ Essen und Trinken ▪ Kleidung ▪ Körper und Gesundheit ▪ Jahres- und Tagesablauf ▪ Feste und Feiern ▪ Kindheit und Erwachsenwerden ▪ Schule und Arbeitswelt ▪ Hobbys und Interessen ▪ Umgang mit Geld ▪ Erlebnisse und Fantasiewelt ▪ Gedanken, Empfindungen und Gefühle ▪ Einstellungen und Werte ▪ Umwelt und Gesellschaft ▪ Kultur, Medien und Literatur ▪ Interkulturelle und landeskundliche Aspekte	▪ (Fictional) Biographies ▪ Simple (technical) descriptions ▪ Diary entries ▪ Directions and instructions ▪ Emails ▪ Letters (personal, advice, application) ▪ Magazine articles ▪ Narrative reports ▪ Lengthy postcards ▪ Statements of personal views and opinions ▪ Stories (create an ending; given an ending – create a story; use a visual impulse to create a story; personal)	▪ To convey emotions, feelings ▪ To inform ▪ To convince, persuade ▪ To entertain, please ▪ To keep in touch ▪ To describe ▪ To give directions and instructions	▪ Self ▪ Others

Table 1: Construct Space for Long Prompts

E8 Construct Space

Prompt Type	CEFR Level	CEFR Descriptor	Deskriptor aus BIST-VO: Schüler/innen können…	Topic Area	Text Types	Intention/ Purpose	Primary Audience
Short Prompt	A2+	▪ Can write about everyday aspects of his/her environment, e.g. people, places, a job or study experience in linked sentences. ▪ Can write very short, basic descriptions of events, past activities and personal experiences.	▪ in Form verbundener Sätze etwas über das alltägliche Umfeld schreiben, wie z. B. über Familie, andere Menschen, Orte, Schule	▪ Familie und Freunde ▪ Wohnen und Umgebung ▪ Essen und Trinken ▪ Kleidung ▪ Körper und Gesundheit ▪ Jahres- und Tagesablauf ▪ Feste und Feiern ▪ Kindheit und Erwachsenwerden ▪ Schule und Arbeitswelt ▪ Hobbys und Interessen ▪ Umgang mit Geld ▪ Erlebnisse und Fantasiewelt ▪ Gedanken, Empfindungen und Gefühle ▪ Einstellungen und Werte ▪ Umwelt und Gesellschaft ▪ Kultur, Medien und Literatur ▪ Interkulturelle und landeskundliche Aspekte	▪ (Fictional) Biographies ▪ Simple (technical) descriptions ▪ Diary entries ▪ Directions and instructions ▪ Emails ▪ Letters (personal, advice, application) ▪ Magazine articles ▪ Notes ▪ Notices ▪ Postcards ▪ Statements of personal views and opinions	▪ To convey emotions, feelings ▪ To inform ▪ To convince, persuade ▪ To entertain, please ▪ To keep in touch ▪ To describe ▪ To give directions and instructions	▪ Self ▪ Others
	A2	▪ Can write short, simple formulaic notes relating to matters in areas of immediate need. ▪ Can write short, simple imaginary biographies and simple poems about people.	▪ kurze, einfache Notizen und Mitteilungen schreiben, die sich auf unmittelbare Bedürfnisse beziehen ▪ einfache Texte z. B. zu Bildimpulsen oder Schlüsselwörtern (key words) schreiben ▪ kurze, einfache Biografien und andere einfache fiktionale Texte schreiben				

Table 2: Construct Space for Short Prompts

5. Structure of the Test

The test contains two sections. Section 1 consists of a short writing task with an expected response of 40 to 70 words. Section 2 consists of a long writing task with an expected response of 120 to 180 words.

The two tasks will be assessed separately on the basis of the four dimensions of the Writing Rating Scale.

6. Time Allocation

Total testing time available: 45 minutes.
Time for administration at the beginning (handing out test booklets): 5 minutes.
Time for administration at the end (word count and collecting test booklets): 5 minutes.
Working time: 35 minutes. The short task should take about 10 minutes, the long task about 20, with 5 minutes for revision.

7. Item Formats

The candidates' scripts will be handwritten on the pages provided in the test booklet. The writing task is guided by prompts that ensure that the candidates produce enough language that makes reliable and valid assessment possible.

The prompts may contain black and white pictures or drawings. They need to be appropriate for the age and at a language level no higher than A2. Input texts should be authentic, if at all possible, and as long as necessary to contextualise the task. Ideally, they should not be longer than 50 words (excluding content points).

The prompts that are developed are to be free of stereotypes. They offer the opportunity to write from experience, but are designed not to intrude on the students' personal feelings.

Prompts need to state the reason for writing, the intended audience and the required text type. The working time available and the number of words for the required length of the texts will be indicated in the instructions.

8. Language Level for Instructions and Prompts

All relevant instructions and prompts are in English with additional information given by the test administrator in German. However, they must be formulated in language that is well within reach of the candidates' expected language level and therefore easily understandable for all test takers. Test takers must not be put at a disadvantage because they have difficulty understanding the instructions or the prompts. The reading competence expected is CEFR level A2.

9. Assessment with Writing Rating Scale

In E8 testing the most significant competences needed for writing are identified for assessment purposes, the first and foremost of which is the communicative competence demonstrated in an appropriate response to the task. In practical terms this means that all expected content points of the prompt are to be clearly and meaningfully mentioned by the test takers. For the higher bands, elaboration of some content points is required. The second is the ability to produce fluent text by using adequate devices to create coherence and cohesion on sentence, paragraph and text level. The

third is a good knowledge of a range of grammatical structures and the ability to use them accurately, and the fourth is the choice of vocabulary that has a certain range, is accurate and relevant to the content.

Therefore the test is designed to elicit language samples that allow the candidates to be assessed in four areas: Task Achievement, Coherence and Cohesion, Grammar, and Vocabulary. The two tasks are assessed separately by trained raters, using an analytic rating scale based on these four dimensions. Multiple-rating and double-rating of a sufficiently large sample of scripts ensure reliability. Differences in rater severity are adjusted for in the process of multi-faceted Rasch analysis.

Pages 25–27 include the four dimensions of the analytic rating scale for writing and more detailed scales for Task Achievement.

Writing Rating Scale (July 2011)

	Task Achievement	Coherence and Cohesion	Grammar	Vocabulary
7	▪ complete task achievement	▪ cohesion on both sentence and paragraph level using a limited number of cohesive devices ▪ clear and coherent text	▪ good range of structures ▪ relatively high degree of grammatical control and few inaccuracies which do not impair communication ▪ message clear	▪ good range of vocabulary communicating clear ideas ▪ generally accurate vocabulary ▪ formulations sometimes varied to avoid repetition
6				
5	▪ good task achievement	▪ good sentence level cohesion as a linear sequence on a simple level ▪ some paragraph level coherence and cohesion ▪ fairly clear and coherent text	▪ generally sufficient range of structures ▪ occasional inaccuracies which can impair communication ▪ message clear	▪ sufficient range of vocabulary communicating clear ideas ▪ occasionally inaccurate vocabulary ▪ major errors possible when expressing more complex ideas
4				
3	▪ sufficient task achievement	▪ some simple sentence level cohesion using simple connectors like 'and', 'but' and 'because' ▪ frequent lack of coherence and cohesion on paragraph level ▪ text often lacks clarity and coherence	▪ limited range of simple structures ▪ frequently inaccurate with basic mistakes, generally without causing break-down of communication ▪ message usually clear	▪ limited range of vocabulary mostly communicating clear ideas ▪ frequently inaccurate vocabulary controlling a narrow lexical repertoire ▪ tendency to use phrases from the prompt
2				
1	▪ some task achievement	▪ basic linear connectors ('and', 'then') on word or word group level ▪ text not coherent	▪ extremely limited range of simple structures ▪ limited control causing frequent breakdown of communication ▪ message seldom clear	▪ extremely limited range of vocabulary communicating few clear ideas ▪ mostly inaccurate vocabulary frequently causing breakdown of communication ▪ several chunks lifted from the prompt
0	▪ no task achievement	▪ not enough assessable language	• not enough assessable language	▪ not enough assessable language

Task Achievement Long Tasks

	Extended Scales
7	▪ complete task achievement with ▪ all content points mentioned & three or more of them elaborated
6	
5	▪ good task achievement with ▪ 85 % of the content points mentioned & two or three elaborated ▪ or all content points mentioned and one or two elaborated
4	
3	▪ sufficient task achievement with ▪ 65 % of the content points mentioned & one or two elaborated ▪ or all content points mentioned without elaboration
2	
1	▪ some task achievement with ▪ 50 % of the content points mentioned & no elaboration
0	▪ no task achievement

Text type requirements:

- It is expected that text type requirements are met. If they are not met (missing/ inappropriate salutation or closing formula; inappropriate register), there is downgrading by one band (two bands if requirements are not met at all).

Text length: 120–180 words

Below-length answers

- Anything below 110 words will be penalised (downgrading by one band).
- Fewer than 80 words - downgrading by two bands (if the general score is band 3 or above; otherwise, downgrade by one band).
- 50–79 words: Assessment is confined to bands 1 and 2.
- Answers containing fewer than 50 words receive 0.

Over-length answers

- More than 180 words: the whole answer is assessed.

Task Achievement Short Tasks

	Extended Scales
7	▪ complete task achievement with ▪ all content points mentioned & one or two of them elaborated
6	
5	▪ good task achievement with ▪ all content points mentioned & possibly one elaborated
4	
3	▪ sufficient task achievement with ▪ 70 % of the content points mentioned & no elaboration ▪ or 50 % of the content points mentioned & one elaborated
2	
1	▪ some task achievement with ▪ 50 % of the content points mentioned & no elaboration
0	▪ no task achievement

Text type requirements:

- It is expected that text type requirements are met. If they are not met (missing/ inappropriate salutation or closing formula; inappropriate register), there is downgrading by one band (two bands if requirements are not met at all).

Text length: 40–70 words

Below-length answers

- Anything below 30 words will be penalised (downgrading by one band).
- Fewer than 30 words: Assessment is confined to bands 1 and 2.
- Answers containing fewer than 20 words receive 0.

Over-length answers

- More than 80 words: the whole answer is assessed.

10. Prompts and Performance Samples with Justifications

The task prototypes below are taken from the 2007 E8 writing test. It is important that the tasks are structured and contain a number of content points so that Task Achievement can be measured.[4]

The instructions below have been reduced to a minimum because most things are announced by the test administrator in German before the test takers open their test booklets. The instructions in English are to ensure that they keep the main points in mind, but also that learners with a first language other than German have the same fair chance to do the task.

10.1 Long Task

Instructions

Read the instructions carefully and then write your text on the next page.
Time: **20 minutes**
Text: **120–180 words**
Use paragraphs.
In your text, try **not** to use language from the task below.

Long Prompt from E8 Writing Test 2007

You have just moved to another town/village. Write a **letter** to your American/English friend in which you tell him/her about your new situation.

Inform him/her about
- your new place of living.
- the reason for moving.

Describe
- the town/village you're living in now (buildings, people,…).
- your new home.

Tell him/her about
- the first days of your 'new life' (new school, teachers,…).
- how you feel about your new situation.

10.1.1 Script 1

Dear Bill,

how are you? Now I'm living in Vienna. That is in Austria. It is a very big city with nice people. There is also a fun fair called "Prater". My parents got divorce and so I'm living here with my mother. There are wonderful buildings in this city like the animal park "Schönbrunn" and many castles. I like the river "Donau" very much, because I often go swimming there. My new house is very big and next to it is a forrest. I like that. The first day of my "new life" was not so good. When I came into my class most of the pupils laughed at me but the teacher was nice. I hope you will write back.

Yours, Raphael

(124 words)

4 More prompts are available on the BIFIE website: http://www.bifie.at/freigegebene-items

Justifications

Task Achievement 5

The text meets the text type requirements, is within the word limit set, and uses an informal register suitable for a letter to a friend. The candidate works his way through the content points, only just touching on the last one. How he feels needs to be inferred from statements like *The first day of my 'new life' was not so good, most of the pupils laughed at me,* and *the teacher was nice.* All the other content points are mentioned. Whereas content point 3 shows good elaboration, the elaboration of content points 1 and 4 is less successful. The passage *Now I'm living in Vienna. That is in Austria.* can be taken as mentioning content point 1 and the following two sentences *(It is a very big city with nice people. There is also a fun fair called "Prater".)* can be seen as an attempt at elaboration.

With content point 4 elaboration is just as thin. The statement *My new house is very big* mentions content point 4, but does very little in the way of describing the new home (like for instance how many rooms, what his/her room looks like, etc.) and the sentence added *(next to it is a forrest. I like that.*) is not really descriptive. So the writer cannot be given much credit for this attempt at elaboration.

This meets the descriptor in the rating scale for band 5: "good task achievement with all content points mentioned and one or two elaborated."

Coherence and Cohesion 4

From the very start the ideas do not connect well. The introductory question *how are you?* is left hanging to be followed by *Now I'm living in Vienna.* The three short sentences that follow add some information about Vienna. But the next idea *(My parents got divorce)* meets the reader unprepared and there is no link to smooth the transition. At this point it becomes clear that the three sentences about Vienna should actually have been moved to content point 3. This way the first unit of the text comprising content points 1 and 2 would have flowed better. Text organisation is based on the sequence of the content points provided with the writer's hand practically invisible.

The text is certainly cohesive at sentence level but hardly at paragraph level. Moving from one idea to the other (telling about the new place – the reason for moving – describing the new town) may be implied by the order of the content points, but the abrupt way this has been done shows the writer's limitations.

However, the text is "fairly clear and coherent". Sentence level cohesion is good and some basic connectors are used to deliver a "linear sequence [of points] on a simple level". There is some paragraph level coherence, but there are no transitions or linking devices between the various ideas presented. While band 5 could be considered for this performance, the lack of paragraph organisation leads to downgrading by one band to band 4.

Grammar 5

At first sight the structures used are generally simple, many of the sentences are very short. There are two identical cases of the present progressive *(I'm living),* four uses of *is, there is* and *there are,* four cases of the present simple *(I like [2x], I often go, I hope),* three simple past tense forms *(was, came, laughed)* and one future *(will write).*

Although all of these tense forms are basic, they seem sufficient to express the writer's ideas and they have been used correctly. Moreover, the writer uses complex sentences correctly, which hints at more complex language competences.

There are actually two subordinate clauses *(because, when)* and a third with the final subordinator *that* omitted in *I hope [THAT] you will write back.* Moreover, if one takes a close look at the other sentences, we find good post-modification of noun phrases *(big city WITH…, fun fair CALLED…, buildings…LIKE…).* Even the co-ordinate clauses have a twist to them *…and SO…, and NEXT TO IT….* So there is definitely sufficient complexity to justify band 5.

Vocabulary 5

The candidate has a sufficient range of vocabulary to express himself. On the one hand, there is rather unexpected vocabulary used (nearly) correctly *(got divorce, fun fair, laugh at)* as evidence of a good range of vocabulary, on the other, there is also repetition of very simple expressions such as *nice, big* and *like,* and there are some occasional inaccuracies *(got divorce, forrest).* The majority of words belong to the most frequently used basic English vocabulary, but the ideas communicated are always clear.

A good range of vocabulary can only rarely be seen, as most parts of the text display a sufficient range. What is definitely lacking is the ability to vary formulations to avoid repetition (band 7), so the text is band 5.

10.1.2 Script 2

Hi Steven!
In Salzburg it is very cool and I'm living in a small flat with 5 rooms two bedrooms a kitchen a livingroom and a bathroom. Our garden is not so big, but big enough for us. The building very beautiful and it give no skyscrapers and it is very hot. I always go in the garden and I lie in the sun. The people are very funny and they accept that I speak english. I have got two new friends and they speak very good english. The teachers are very good and we have a lot of english and I'm the best one, but in Deutsch I'm very bad. I feel very good with my new situation and I wish all my old friends and the teachers a good luck for the next time and I hope you always wish me a good next time.
Yours Olav! *(149 words)*

Justifications

Task Achievement 4

The text mentions all content points with the exception of content point 2, where reasons for moving should be stated (=85%). There is some elaboration of content point 4 by describing the size of the garden and content point 5 by supplying the reader with some additional information about his new friends *(they speak very good english)* and his achievements at school (lots of English lessons and him doing well in English), although this only implicitly refers to the content point "the first days of your 'new life'". Content point 6 has been lifted from the prompt adding the word *good.*

As content points 1 and 6 have only been dealt with in an extremely basic way by merely mentioning Salzburg and lifting a phrase from the prompt and the elaboration of content points 4 and 5 is rather weak, a downgrade to band 4 is the consequence.

Coherence and Cohesion 2

Text organisation is quite low with simple addition as the dominating structuring principle. The simplest connector *and* occurs with undue frequency, proving that ideas are mostly strung together without expressing logical relations. Apart from the second sentence *(Our garden is not so big, but big enough for us)* the text does not read well because some sentences that follow each other have little or no connection at content level, so the text lacks coherence. Some chunks of language that have little in common are often joined in one sentence, e.g. *The building very beautiful and it give no skyscrapers and it is very hot.*

Therefore, this text is characterised by a noticeable lack of clarity and coherence and some rather basic sentence level cohesion. This would point towards a weak band 3. As there are no paragraphs, the text has to be downgraded to band 2.

Grammar 3

The writer uses a quite limited range of simple structures correctly, repeating the same basic pattern with little variation. He only uses present tense structures, although content points 2 and 5 would invite the use of past tense. Therefore, the range of structures cannot be considered sufficient for the purposes of the task. Even within the narrow frame of present tense sentence structures there is inappropriate use of the continuous form in *I'm living in a small flat.* The simple message is usually clear, although an error such as *it give no skyscrapers* causes breakdown of communication. Similarly, the phrase *the building very beautiful* leaves us undecided whether it should refer to the house where the writer lives or the buildings in Salzburg. All this would suggest a very weak band 3, which, however, is supported by the relatively high degree of correctness.

Vocabulary 4

The text shows successful control of a limited range of vocabulary, with some good phrases sticking out such as *big enough for us, skyscrapers,* or *they accept that I speak english.* The simple vocabulary used in the first part of the text communicates mostly clear ideas, but in the last sentence the writer seems to be attempting too much, leaving the safe area, and this results in several breakdowns *(I wish them a good luck;* L1: *for the next time; I hope you always wish me a good next time)* demonstrating the limitations, as does the use of *Deutsch* for the subject German. The narrow lexical repertoire and also the tendency to lift phrases from the prompt *(I feel very good with my new situation)* would indicate a band 3, but the occasional neat expression and the fact that severe problems appear only when trying to express a more complex train of thought (tolerated at band 5) justify a weak band 4.

10.2 Short Task

Instructions

Read the instructions carefully and then write your text on the next page.
Time: **10 minutes**
Text: **40–70 words**
In your text, try **not** to use language from the task below.

Short prompt from E8 Writing Test 2007

Your friend's birthday party was a few days ago. Write an **email** to tell him/her that you liked the party.

- **Tell** him/her why you liked the party.
- **Tell** him/her what you liked best.
- **Ask** your friend when you are going to meet again.
- **Suggest** something for the next weekend.

10.2.1 Script 3

Dear Daisy, how are you? Let's talk about your party. It was so great! I liked the party best, because Lukas was there. And I liked the games we played. Oh and tell you mum, that the food was excellent! What are you going to do on Sunday? Maybe we can go to cinema or swimming. Tell me please if you have time. I nearly forgot it: Tanja's birthday party is in two weeks, she invited me, are you invited too? Okay I have to help my mum with dinner.
Love you big kiss
Yours,
Aida *(95 words)*

Justifications

Task Achievement 7

The register and the layout are clear indications that this text is an email. The salutation and closing formulas are most appropriate and for these reasons text type requirements are perfectly met. All content points have been mentioned and there is elaboration of content point 1 *(It was so great!, I liked the games we played, the food was excellent)* and content point 3 as the candidate makes enquiries about an upcoming event in the near future *(I nearly forgot it: Tanja's birthday party is in two weeks, she invited me, are you invited too?),* so we have complete task achievement – band 7.

Coherence and Cohesion 7

The text admirably incorporates qualities of spoken English (*Let's talk about, Oh and, I nearly forgot, Okay I have to),* which one would expect in an informal email, and which make it flow well. A number of cohesive devices are used to connect groups of sentences together very well, such as lexical cohesion (*party-it-party),* conjunctions *(because-and),* backward and forward referencing (*we played-tell your mum that),* and there is evidence of good linear sequencing of points making it a clear and coherent text, pointing it towards band 7. However, there are two abrupt changes in the linear sequence of the text *What are you going to do* and *I nearly forgot,* but as paragraphs are not expected in short texts, it remains a band 7.

Grammar 7

There is a relatively high degree of grammatical control with only one slight slip in accuracy. However, the omission of the definite article in the phrase *go to cinema* does not impair communication and the message throughout the text is clear, suggesting band 7. The candidate's good use of the present, past, going-to-future, Saxon genitive, and a subordinate clause proves he/she is able to address all the language functions included in the prompt: to inform, to ask (*how are you?, What are you going to do on Sunday? are you invited too?),* and to suggest *(Maybe we can go to cinema).* This indicates the good range of structures the candidate is able to use accurately, thus supporting a strong band 7. The very casual sentence *Okay I have to help my mum with dinner* with near-native omission of the definite article *(with dinner)*, is further evidence that clearly points to a band 7.

Vocabulary 7

The vocabulary elicited by the prompt points towards band 7. Not only does it contain a good variety of appropriate and accurate content words *excellent, invite, dinner,* but also many collocations that are equally appropriate and accurate and lend a certain naturalness to the text (*the food was excellent; what...going to do on Sunday; have time; in two weeks; I nearly forgot; love you big kiss).* Another indicator for band 7 is the candidate's choice of words which enable her to get her message across very clearly throughout the text. Some of the phrases *(Let's talk about; Oh and tell; I nearly forgot; Okay I have to)* indicate a certain air of 'chattiness' to the text. Furthermore, the use of *Love you big kiss,* as an alternative or addition to the common closing line *Yours,* which exemplifies how the candidate can vary formulations to avoid repetition, is additional evidence that this is a band 7.

10.2.2 Script 4

Hey Kevin!!! I've liked your party because it was very cool. What I liked most were all the nice girls and all the nice waterpipes. Can we do such a party again tomorrow? It would be very nice. But please, buy more waterpipes, and more grass! And next weekend, we can do it again, or? I have a better idea, we can go to the city and chillout at a concert.
(65 words)

Justifications

Task Achievement 4

The text clearly follows the organisation of the content points *(liked the party because it was cool/liked the girls and waterpipes best, do such a party tomorrow again?, go to the city …).* Although the first content point is handled very briefly and content point 3 features as an indirect request one can still say that all content points have been mentioned, which hints towards band 5. Bands 6 or 7 cannot be taken into consideration because there is no elaboration. Although content point 4 consists of two sentences, the introductory question to content point 4 is nothing else than a repetition of the question in content point 2 *(Can we do such a party again…?, … we can do it again, or?)* and the actual content point 4 is covered by the following suggestion *(I have a better idea …).*

As the closing formula is missing, text type requirements are only partly met (salutation: *Hey Kevin!!).* This leads to downgrading by one band to band 4.

Coherence and Cohesion 5

The first part of the text (sentences 1–5) shows good sentence level cohesion through backward and forward referencing, thematic fronting and lexical repetition *(party-it; what I liked most; such a party; it would be ...; waterpipes).* This part, though not marked as such in the layout, even shows some good paragraph level coherence within the narrow scope of a short text.
The concluding part *(And next weekend....)* provides more back referencing in *it* for *party,* and the following suggestion is thematically linked to the previous question through *I have a better idea.*

Therefore, we can ascertain "good sentence level cohesion" and "some good paragraph level cohesion", which justifies band 5.

Grammar 5

The text clearly shows a sufficient range of structures to fulfil the task, which requires the description of a past event, the question for a future event, and a suggestion. The past event is correctly described *(was very cool; what I liked most were),* the question is correctly phrased and followed by two statements making correct use of a subjunctive *(would)* and a polite imperative *(But please, buy more…).* The text also features a comparative, the correct use of a modal verb form, and a subordinate clause.

The text is generally very accurate; the mistakes in describing a past event *(I've ~~ve~~ liked)* and the L1 interference in the use of *or?* instead of a tag question do not impair communication and it is clear what the writer wants to say. Hence, band 5 is appropriate.

Vocabulary 4

The candidate shows sufficient lexical range to fulfil the task and communicate his ideas, which would hint towards band 5. However, in doing so he makes use of very simple vocabulary: *very cool, all the nice girls, go to the city, concert, was/were, buy, more, again* etc. The only words sticking out are *waterpipes, grass, such a* and *to chill out.* Although the words *water* and *pipes* have been combined and are semantically 'new', the words themselves are still extremely basic, as is *grass.*

This leaves us with as little as *to chill out* and *such a* as the only lexical items that would go beyond a limited range.

On the other hand, vocabulary is accurate with *to ~~do~~ a party* as the only incorrect use, which does not create any misunderstanding. Taking into consideration that the writer does not take any lexical risks and the items used are simple, limited in range and repetitive a downgrading from band 5 to band 4 is appropriate.

10.2.3 Script 5

My friend have at 7.5. birthday. The birthday-party was very good. We had a lot of fun on the party. We play playstation and we went play football but the best was that we are ate pizza. We go at the weekend to a football match.
(46 words)

Justifications

Task Achievement 0[5]

The candidate tries to work his/her way through (most of) the content points. Content point 1 is mentioned as the student writes that *the party was very good* and that they *had a lot of fun.* Then he/she mentions a few things that they did at the party. It can be assumed that these activities are the reasons why he/she liked the party. Content point 2 is short but ok *(the best was that we are ate pizza).* Content point 3 is missing completely, and content point 4 is not recognisable as a suggestion, therefore not mentioned correctly *(We go at the weekend to a football match).*

So there is some task achievement with 50 % of the content points mentioned but no elaboration, which points towards band 1.

However, the text does not meet text type requirements at all. The candidate was asked to write an email to a friend. The first thing that is striking here is that there is no salutation and no closing formula. All in all, this text reads like a report rather than an email message. As a consequence, the text has to be downgraded by two bands, which means it is to be placed at band 0.

Coherence and Cohesion 2

At first sight coherence and cohesion in this text seem to be rather poor. However, there is some simple sentence level cohesion connecting the ideas of content points 1 and 2 (friend's birthday – the birthday party – fun at the party – activities at the party). We find one instance of a cohesive definite article *(the* party) and the logical connection between *We had a lot of fun on the party. We play playstation and [...]* is actually fairly obvious.
Moreover, in one sentence some basic connectors are used fairly successfully *(and* in order to link two main clauses; correct use of *but + that-clause).*

The last sentence breaks off the coherence that was there and disrupts the minimal quality of flow that the first part of the text has.
That is why coherence and cohesion is clearly better than band 1, but not good enough for band 3, hence (a weak) band 2.

Grammar 1

The text shows an extremely limited range of simple structures. Most sentences follow a very basic subject-predicate-object pattern.

The task basically requires the description of a past event, the question for a future event, and a suggestion, so altogether rather simple structures. However, even some simple structures are used incorrectly *(we went play football; we are ate pizza).* What is more, there is no question, no phrase that indicates a suggestion, and even the very basic verb forms to indicate a present, past or future time aspect are quite often used incorrectly *(my friend have; we play playstation* for describing a past event; *We go at the weekend to a football match).*

5 As band zero for Task Achievement has not been reached directly, but after downgrading, the other dimensions are assessed without any restrictions. Script 6 demonstrates that when Task Achievement is a clear zero from the outset, the other dimensions are not rated as there is not enough assessable language.

In addition, there are examples of the misuse of prepositions in very basic phrases *(on the party; at 7.5.),* and wrong word order in very basic sentence structures *(My friend have at 7.5. birthday; We go at the weekend to a football match).*

Although there is no breakdown of communication and the message is usually clear, the extremely limited range and incorrect use of simple structures make this text a performance at band 1.

Vocabulary 1

The text shows an extremely limited range of vocabulary, using very basic words, e.g. *friend, birthday, birthday party* (mentioned in the prompt), *a lot of fun, football match, pizza,* and forms of *have, be, play* and *go.*

Although the candidate is able to communicate his/her very simple ideas successfully and (mostly) accurately, the extremely limited range of vocabulary overrules accuracy and this makes the performance band 1.

10.2.4 Script 6

My friend's birthday party was a few days ago. Write an email to tell ihm that you liked the party. Tell him why you liked the party. Tell him what you liked best. Ask your friend when you are going to meet again.
Suggest something for the next weekend.
My best friend's. Name von my best friend's is …
(58 words)

Justifications

Task Achievement 0

There obviously is no task achievement as the candidate has merely copied the given input text and the content points instead of doing what they said. The only addition made by the writer is minimal and bears no relation to the task set.

On the grounds of task achievement being band zero, the other three dimensions are not assessable as we could only assess the language used in the task description, but not any of the candidate's competences.

Coherence and Cohesion 0

No enough assessable language in terms of coherence and cohesion.

Grammar 0

Too little independently produced language to allow assessment.

Vocabulary 0

Too little independently produced language to allow assessment.

Scale Interpretations

In the rater training courses over the past three years it has become clear that the rating scale as it stands is by no means self-explanatory and ready for general use. Therefore, some comments on how to read and interpret the scale are added here.

Scale Interpretation – Task Achievement

The scale on Task Achievement has no direct correlation with the CEFR and assesses the content components of a text and text type requirements. The way the descriptors are formulated leaves room for differing interpretations, which will lead to diverging assessments. In order to reach the aims of the training course and the standardisation meetings, a common understanding of what all the elements in the scales mean and how they relate to each other is required. To improve inter-rater reliability we need to clarify the key terms.

CONTENT
The first issue that seems simple enough is to decide whether a content point has been mentioned in a script or not. Whereas this is quite straightforward in most cases, there is room for confusion when, for instance, some key words from the prompt appear in the text, but the language around them does not make much sense. As it says in the test specifications that "all expected content points of the prompt are to be **clearly and meaningfully** mentioned" (p. 23), such a content point would not be considered as being mentioned.

What rater trainees have found most challenging and confusing when assessing Task Achievement is distinguishing between **mentioning** a content point and **elaborating** it. With reference to the long prompt on p. 28 the following example, taken from one of the scripts used in the training sessions, can serve to demonstrate what "mentioning a content point" means: *We live now in New York near the Central Park, we moved because my mum had not found a job.* In this sentence, content points 1 and 2 are mentioned. The text goes on: *But in New York my mum has a good job.* This sentence extends point 2 a little, but as it is little more than a reformulation of the previous sentence, it cannot be seen as elaboration.

Elaboration of this point could have been something like this: *In New York she works as a secretary in a bank on the 35th floor of a high building in Manhattan and is quite happy.* Or: *In New York she sells pancakes in the streets, and she is happy.* Good elaboration then involves the introduction of a new idea, a real extension of what has been said before.

Less successful elaboration is something you recognise when you see it. Look at this example – Content point 5 of the long prompt on p. 28: "Tell him/her about the first days of your 'new life' (new school, teachers, …)": *I have new friends but you are forever my best friend. The new school is very big. The teachers are sometimes unfriendly. And the school colleagues are not polite.*

The first sentence mentions new and old friends. Then it moves abruptly to the cues from the prompt and adds some simple words to each. Finally, there is a new sentence based on the same pattern. – There is some elaboration here, no doubt, but it is not very good. This will be reflected in the assessment.

Whereas on one level we can assess Task Achievement quantitatively by simply counting the content points mentioned and elaborated respectively, the discussion

above makes it clear that, in addition to this, there is also a qualitative component to be considered. The first question is *How many?,* but the second is *How good?.*

TEXTS AT BAND ZERO
In the following cases there is no assessment in any of the four dimensions:

- Texts that do not deal with the given topic and the content points listed.
- Texts that are extremely rude, sexist, racist, or propagating violence.
- Texts that show an attempt at dealing with the topic but do not contain enough assessable language, i.e. fewer than 50 words in long texts and fewer than 20 words in short texts.

Texts that are placed at band zero due to downgrading (text type requirements and/or text length) are assessed in the other three dimensions.

TEXT TYPE REQUIREMENTS
Another term in the Task Achievement scale that has invited frequent discussions in the training sessions is **text type requirements.** The issue, however, is greatly simplified by the context of use, which is tested at E8 level. This simply means that in the given test situation there is little room for stylistic variation on the part of the test takers. In most cases an informal register is the only one they have access to, and test takers are not expected to introduce stylistic differences related to particular text types like magazine articles, reports, diaries, letters, or emails. Test takers will be using a more or less informal style in all their texts and the only thing they need to know is how to open and close a letter or an email. In a rater meeting it has been decided to consider emails as slightly more informal than letters, but still requiring some kind of salutation and closing formula.

In practical terms – and in the context of E8 testing – this means that meeting of text type requirements is considered a given requirement so the test takers do not get any bonus for it. It is only in case of problems in this area that we take this into consideration and react by downgrading one band or two bands respectively. The following guidelines have been discussed and agreed on in past rater meetings:

- Salutation AND/OR closing formula missing or wrong – downgrade by one band.
- Serious register/style problems – downgrade by one band.

TEXT LENGTH
Text length is a related issue, which has been set down in sufficient detail as a footnote in the Task Achievement scales. The main point is that over-length texts are not penalised whereas texts that are significantly below the requested number of words (110/80 and 30 respectively) are downgraded. This is based on the assumption that a writer who only delivers two thirds or less of the length required will have serious problems to produce a substantial text.

The baseline testing of 2009 has shown that texts within the range of 80–109 words rarely get a provisional score that is higher than band 4, those within the range of 50–79 words get no higher scores than band 3. At these low performance levels, however, a particular problem has arisen. If, for instance, a "long text" of only 78 words is provisionally placed at band 2, it will be downgraded by two bands for text length, so the final score would be zero. In a case like this (with the word count so close) it has been decided to downgrade only by one band for text length so as to recognise overall task achievement by placement in the lowest band.

Scale Interpretation – Coherence and Cohesion

Coherence is a quality criterion that refers to the logical arrangement of ideas and arguments within a text. In a well-written – that is "coherent" – text the writer successfully arranges his/her sentences to achieve a purpose, e.g. to reflect the chronological sequence of events or to develop a convincing line of argument. A coherent text makes it easy for the reader to follow the writer's train of thought so that there is no need to stop and reread in order to establish meaning as ideas and arguments flow smoothly and logically. A less coherent text, however, impairs readability and appears jumpy.

The term **cohesion** relates to the relationships between elements of a text, which is the way words, word groups and individual sentences are linked. There are several ways in which cohesion can be established. Simple sentences can be connected by using linking words such as *and, but* or *because.* For example, the sequence *My holiday was a disaster. It rained almost every day.* can be reformulated as *My holiday was a disaster because it rained almost every day.* Another solution would be keeping the two sentences but linking them by saying *It rained almost every day. Therefore my holiday was a disaster.* Some such cohesive devices that we may expect writers to use at level A2/B1 are:

Addition	and, or, also
Time	when, after, before,
Result	so, therefore
Contrast	on the one hand – on the other hand, although
Reason	because, as
Exemplification	for example
Sequence	first, then, next, finally

We will, however, have to bear in mind that a text can be coherent even if very few of these cohesive devices are used and that, on the other hand, the frequent use of cohesive devices does not necessarily turn an incoherent text into a coherent one.

Other techniques to make a text appear cohesive are references by the use of personal pronouns, possessives, demonstratives, and comparatives. At a very simple level, in the two sentences *My best friend is Michael. He is in the same class as I.* cohesion is realised by using the personal pronoun *'he'* instead of repeating the name *Michael.* Similarly, in a passage such as *My sister has the big room in the house. Mine is a lot smaller.* the possessive *mine* refers to the room in the previous sentence, thus linking the two sentences. Demonstratives can serve the same purpose. In *I got a new camera for my birthday. That was my best present ever.* the word *that* refers to the camera the writer got, thereby linking the two sentences successfully.

Sentences can also be connected by substituting one or more words in a sentence. In *We have a lot of field trips in our school. The nicest one was to Schönbrunn Zoo.* the writer has replaced *field trips* by *one* in the immediately following sentence. In *Girls are better at English. Everybody thinks so.* the word *so* represents the whole idea that girls are doing better at languages. A particularly successful way of establishing cohesion in a text is the use of lexical chains as exemplified in the following text passage:

When I think of clothing I would say that T-shirts with crazy designs like dots, squares, skulls are definitely in. All my friends are wearing that and they think it's the latest craze! This year wearing the 'right' shoes like 'Converse' or 'Vans' is very important. Everybody loves to wear them because it's a must-have!

In the first sentence the writer uses the phrase *definitely in* to describe certain kinds of fashionable T-shirts. In the following sentence this idea is continued by using the phrase *the latest craze.* At the end of the paragraph this concept of a fashion product is reformulated as *it's a must-have.* This establishes a lexical chain that binds the sentences together and establishes a smooth flow of ideas in the paragraph.

In the CEFR, coherence and cohesion is an aspect of the pragmatic competences of a language user. The discourse competences relevant for writing that are dealt with in the CEFR are "flexibility to circumstances", "thematic development" and "coherence and cohesion". As the latter is the quality criterion that is of particular relevance within the range of A2/early B1 writers, it is the one that is represented in the E8 Assessment Scale. In the E8 Scale for coherence and cohesion the CEFR (see p. 125) moves from the very basic A1 skill of being able to link words with linear connectors such as *and* or *then* to A2, which means also successfully using connectors that express reason *(because)* and contrast *(but)* to link words or word groups. A2+ includes the ability to use these most frequent connectors to describe something as a "list of points", whereas one level further up at B1 the loosely "connected list of points has become a fully connected linear sequence of points".

Our E8 Assessment Scale for writing includes the aspects of both coherence and cohesion. Regarding coherence we expect a text to be essentially clear in its message and coherent at bands 5 to 7, but accept some amount of vagueness and ambiguity in band 5. Band 3 texts are characterised by frequently incoherent text elements, noticeably impairing clarity and readability, while band 1 texts are not coherent at all and consist of mostly disconnected chunks of language. In such band 1 texts we only find the most basic linear connectors such as *and* or *then* as cohesive devices on word group level, while band 3 texts should already show simple sentence level cohesion with a wider range of connectors. A band 3 writer is able to link sentences successfully using simple connectors, but usually fails to produce longer stretches of connected language at paragraph level, making a text appear as a choppy list of points rather than a longer connected sequence. From band 5 up we can demand this longer connected sequence of sentences, with the writer being able to link sentences into clear paragraphs. At band 5 we want to see this ability reflected in at least some parts of the texts, while at band 7 the whole of the text should reflect good sentence as well as paragraph level cohesion. At the top band 7 an expert writer will probably not only manage to link sentences smoothly and logically to produce a coherent paragraph, but might also already establish links from paragraph to paragraph. Needless to say, the degree in which handling such issues of coherence and cohesion can be mastered by a test taker also depends on the complexity of the ideas put forward and may have to be taken into account by the rater. The more complex and unexpected ideas there are in a text, the more we have to accept some jumpiness in the way they are presented.

The content points in the prompt will already suggest a paragraph organisation to the writer, but it is finally the decision of the test taker how he/she chooses to organise his/her text. The ability to structure a text of around 150 words into meaningful paragraphs is an important skill that we expect test takers to demonstrate in the E8 Writing Test. In long texts we expect paragraphs from band 3 up, and a lack of indentation or visual marking of paragraphs will result in downgrading the text by one band. A sequence of individual sentences marked as a paragraph cannot be accepted as successful paragraphing if the sentences are arranged in a haphazard and random way showing no connection whatsoever. The same applies to paragraphs consisting of one sentence only. In short texts coherence and cohesion is generally more difficult to demonstrate. Paragraphs are not mandatory and, if used to good effect, could be considered a reason for upgrading the text.

Scale Interpretation – Grammar

The Scale for Grammar comprises descriptors for range, control, and the clarity of the message. Therefore, the raters evaluate the test takers' ability to make use of a range of grammatical structures, the level of their accuracy as well as their impact on the message. The focus is on grammatical forms that create meaning and that are reasonably correct to accomplish successful communication.

Short tasks are designed to be A2 tasks and the range of grammatical structures that is likely to be elicited in such tasks comprises structures typically mastered at A2 level.[6] Long tasks should have the potential to produce B1 language and, as a consequence, also grammatical structures representative of B1 level.[7]

THE CONCEPT OF GRAMMATICAL RANGE
Grammatical range refers to the variety of grammatical structures found in a performance. Range can surface in the variety of grammatical forms (verb modification, tense, aspect, comparative forms, superlative forms) and the complexity of sentences (main clauses, subordinate clauses, conditional or relative clauses) used in a text.

In the E8 context, grammatical range must be seen in relation to the task. We cannot expect the test takers to use structures that are not meaningfully elicited by the task. Since the writing prompts focus exclusively on familiar topics and have to cater for all ability levels, they are as straightforward in their set-up as possible. This does not automatically suggest that the response cannot be more complex than the stimulus. Even if a task is simple in nature, we expect differentiation in grammatical forms or clause types, such as conditional or relative clauses.

Verbs, for example, can be modified, mark aspect, and determine various types of sentence function such as statement, question, negation, command, and exclamation. Moreover, they can be used in their active or passive forms, and test takers may choose to use direct or indirect speech.

In addition to the specifications of the prompt, which will try to elicit certain grammatical structures for task fulfilment, the time allotment and the expected number of words will also have an impact on range. That is, short tasks are likely to provide fewer opportunities to show grammatical range than long tasks.

RANGE VERSUS ACCURACY
In a mistakes and correction driven tradition of teaching, the use of grammatically challenging language can become a problem for a learner if errors occur. Not so in E8 testing. It is E8 testing policy that range overrules accuracy in the sense that rich grammatical range through risk taking is encouraged, while minor inaccuracies that do not impair meaning play a reduced role. The more varied the grammatical range, the higher the band. Risk taking, which results in rich structures but reduced control, can even be a reason for upgrading a text.

Global errors, i.e. errors that interfere with the comprehensibility of the text, will cause downgrading or the placement of a text at a low band. Local errors which do not hinder communication will not automatically lead to downgrading unless their frequency impairs the message or the readability of the text.

6 For an inventory of grammatical areas at A2 level see KET Handbook, 8–9. Available at: http://www.cambridgeesol.org/assets/pdf/resources/teacher/ket_handbook.pdf [24 June, 2011]
7 For an inventory of grammatical areas at B1 level see PET Handbook, 7–8. Available at: http://www.cambridgeesol.org/assets/pdf/resources/teacher/pet_handbook.pdf [24 June, 2011]

Test takers are encouraged to make use of their full potential and the more creative the structural features they show, the better. Nevertheless, the use of variation should not be exaggerated either. The tasks suggest certain scenarios which require special structural solutions. These should produce authentic and natural variation but not artificial texts.

ASSESSMENT OF PERFORMANCES
The placement of a performance at a certain band reflects the range of grammatical structures and the level of their correctness within a meaningfully and successfully accomplished communicative task.

Band 7 texts feature good grammatical range which creates meaning and natural language within the framework of the task. The writer varies the grammatical structures the prompt elicits and may occasionally go beyond the obvious and expected. However, any enhancement should not make the text sound unnatural or result in exaggeration of grammatical structures (range for the sake of range). In addition to good range a relatively high degree of grammatical control is expected in band 7 texts. A few inaccuracies can occur but they will not impair communication.

Band 5 texts show sufficient range of grammatical structures. Sufficient range is achieved, if the writer makes enough use of the prompt's structural features to make the required communication successful and if the grammatical forms used create appropriate meaning. Occasional inaccuracies which can impair communication can be tolerated.

Band 3 texts feature a limited range of simple grammatical structures. This means that the grammatical structures are just enough to achieve successful communication. Mostly they are very simple, repetitive, and hardly varied. Grammatical structures in band 3 texts can be frequently inaccurate and may show basic mistakes. Generally, these mistakes do not cause breakdown of communication.

Band 1 texts feature an extremely limited range of simple structures. This usually forces the writer to compromise the message regarding meaning, content, and naturalness of language. Extremely limited range results in structures that are repetitive and in very simple subject-predicate-object sentence patterns. The structures used hardly go beyond the learnt repertoire of beginners. In addition to structural restrictions, band 1 texts show limited control which frequently causes breakdown of communication.

Scale Interpretation – Vocabulary

When we assess vocabulary we are looking at content words (nouns, full verbs, adjectives, adverbs), collocations and chunks of language that a writer uses to perform a written communicative task. What we need to assess is lexis creating meaning that is reasonably correct to accomplish successful communication. Similar to the grammar scale, the scale for vocabulary also comprises descriptors for range and control.

THE CONCEPT OF LEXICAL RANGE
Range refers to the breadth of vocabulary a candidate uses in a written text. In the E8 context, range must be interpreted in relation to the prompt as raters can assess only the vocabulary actually elicited by the prompt. The time allotment and the expected text length have an impact on range. Short tasks are likely to provide fewer opportunities to demonstrate vocabulary range than long tasks. As mentioned in the previous chapter, short tasks have been designed to be A2 tasks and long tasks have been designed to be B1 tasks. For these reasons the range of lexical items that we can expect in short tasks are words and phrases typically mastered at A2 level[8], for long tasks we can expect a fair amount of words and phrases typically mastered at B1 level[9].

Even if a task is simple in nature we may expect differentiation within choice of words. For example, if a task asks for a narrative description about the first few days at a new school, the texts will primarily contain words related to school, teachers, subjects, new friends etc., which, however, can be varied and modified. Although the prompt language is as simple as possible, writers may well produce a response that exceeds the prompt stimulus.

RANGE VERSUS ACCURACY
It is not enough for a candidate to use a large number of different words in a text to achieve a high band in assessment. The words a candidate chooses must be relevant and appropriate to the topic and used in a way that the candidate is able to convey his/her ideas meaningfully. A top writer among our test takers will use vocabulary that is generally accurate enough to formulate even a more complex idea with clarity. Test takers who stay in absolutely safe language areas (e.g. language picked up in years one and two) and avoid taking any risk have less evidence of mistakes. However, it is E8 policy to encourage our candidates to venture out of their safe language zone by rewarding risk taking to communicate successfully.

Texts that show a good range of vocabulary at band 7 contain a good selection of content words and phrases that demonstrate that the candidate is able to express him/herself clearly and precisely and occasionally can even vary formulations so as not to appear repetitive. We may well expect one or the other expression to stick out and exceed what we typically expect from test takers at this level.

Band 5 texts contain a sufficient range of mostly high-frequency words that again meet the need to communicate clear ideas and are generally used accurately. There may be some occasional mistakes, particularly when the candidate is trying to communicate a more complex idea.

In a band 3 text we expect the lexical range to be limited, containing only a rather narrow repertoire of high-frequency words, but still the simple ideas that are com-

8 Available at: https://www.teachers.cambridgeesol.org/ts/digitalAssets/113295_ket_vocablist09.pdf [24 June, 2011].
9 Available at: https://www.teachers.cambridgeesol.org/ts/digitalAssets/113298_pet_vocablist09.pdf [24 June, 2011].

municated are mostly understandable even if there is a certain amount of inaccurate vocabulary. With band 3 candidates we are likely to detect examples of lifting phrases from the prompt to compensate for their lexical limitations.

Finally, in a band 1 text a writer with extremely limited lexical competence in English will demonstrate this by including only a few very high-frequency content words which are more often than not inaccurate and inappropriate. We commonly expect band 1 writers to compensate for their lack in lexical range by heavily lifting directly from the prompt and by interspersing their text with L1 words in order to 'keep going', thus having the 'knock on effect' of frequently causing breakdown in communication.

The nature of some prompts makes it almost impossible to avoid lifting and raters must take care not to fall into the trap of automatically placing a text at band 4 or below because there is evidence of prompt lifting. A good writer does not just 'copy and paste' words or phrases, but can adapt them and incorporate them successfully into the text to accomplish the communicative task. This is a skill that needs to be acknowledged positively.

An aspect of lexical accuracy that raters need to address is spelling. It is common practice amongst teachers to take marks off for incorrect spelling. However, the emphasis on communicating meaning successfully is central to the E8 context. A text containing many spelling mistakes, in particular those mistakes that change the whole meaning of a word, is very likely to disturb the reader and cause a breakdown of communication. Raters need to assess the extent of breakdown and rate as is necessary. However, as we encourage our writers to take risks, slight 'slips of the hand' and minor errors in spelling that do not change the meaning of the word (e.g. *pleas, tomorow)* should not be penalised. In the end it is the lexical range that is more important than accuracy and a text might merit one of the higher bands despite the inaccuracies.

Literature

Alderson, J. C. 2004. Washback in language testing. Research contexts and methods. In: Cheng, L., Watanabe, Y. & Curtis, A. (Eds.). *Context and Method in Washback Research: The influence of language testing on teaching and learning.* Mahwah, NJ: Lawrence Erlbaum.

Alderson, J. C., Clapham, C. & Wall, D. 1995. *Language Test Construction and Evaluation.* Cambridge: University Press.

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing.* Oxford: University Press.

BIST-Verordnung. 2009. See *Verordnung der Bundesministerin.*

Breit, S. & Schreiner, C. (Eds.) 2010. *Bildungsstandards: Baseline 2009 (8. Schulstufe). Technischer Bericht.* Salzburg: BIFIE. Available as download from http://www.bifie.at/buch/1056 [14. April, 2011]

Brock, R. & Haslinger, U. (Eds.) 2011. *Bildungsstandards für Fremdsprachen (Englisch) 8. Schulstufe. Praxishandbuch* (Neubearbeitung 2011). ÖSZ Praxisreihe 4. Bifie Wien.

Canale, M. & Swain, M. 1980. Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. In: *Applied Linguistics* 1 (1), 1–47.

Council of Europe (Ed.). 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: University Press.

Ek, J. A. van [Council of Europe]. 1979. *The Threshold Level for modern language learning in schools.* Harlow: Longman.

Ek, J. A. van & Alexander, L. G. 1980. *Waystage English.* An intermediary objective below Threshold Level in a European unit/credit system of modern language learning by adults. Prepared for the Council of Europe in association with M. A. Fitzpatrick. Oxford: Pergamon Press.

Field, J. 2004. *Psycholinguistics. The Key Concepts.* London: Routledge.

Field, J. 2005. *Second language writing: a language problem or a writing problem?* Paper presented at IATEFL Research SIG 'Writing Revisited' conference, Cambridge, 25–27 February 2005.

Gassner, O., Mewald, C. & Sigott, G. 2007. *Testing Reading. Specifications for the E8 Standards Reading Tests. LTC Technical Report 2.* Wien: bm:ukk. Available as download from: http://www.bifie.at/publist-07-05-14 [24 June, 2011]

Grabe, W. & Kaplan, R.B. 1996. *Theory and Practice of Writing. An Applied Linguistic Perspective.* London: Longman.

Hyland, K. 2002. *Teaching and Researching Writing.* London: Longman.

Kellogg, R. T. 1994. *The Psychology of Writing.* New York: Oxford University Press.

Kellogg, R. T. 1996. A model of working memory in writing. In: Levy, C. M. & Rands-dell, S. (Eds.) *The Science of Writing.* Mahwah, NJ: Lawrence Erlbaum, 57–72.

KET Handbook. Cambridge UCLES. 2009. Available as download from http://www.cambridgeesol.org/assets/pdf/resources/teacher/ket_handbook.pdf [14 April, 2011]

KET Vocabulary List. Cambridge UCLES. 2006. Available as download from https://www.teachers.cambridgeesol.org/ts/digitalAssets/113295_ket_vocablist09.pdf [14 April, 2011]

Lehrplan der Hauptschule: Lebende Fremdsprache (alle Sprachen) 2008. Available as download from http://www.bmukk.gv.at/medienpool/16682/bgbl_nr_ii_210_2008.pdf [14 April, 2011]

Lehrplan der AHS: Lebende Fremdsprache (alle Sprachen) 2006. Available as download from http://www.bmukk.gv.at/medienpool/782/ahs8.pdf [14 April, 2011]

Lumley, T. 2005. *Assessing Second Language Writing. The Rater's Perspective.* Frankfurt: Peter Lang.

Papajohn, D. 1999. The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. In: *Language Testing* 16 (1), 52–81.

PET Handbook. Cambridge UCLES, 2009. Available as download from http://www.cambridgeesol.org/assets/pdf/resources/teacher/pet_handbook.pdf [14 April, 2011]

PET Vocabulary List. Cambridge UCLES. 2009. Available as download from https://www.teachers.cambridgeesol.org/ts/digitalAssets/113298_pet_vocablist09.pdf [14 April, 2011]

Read, J. 1990. Providing Relevant Content in an EAP Writing Test. In: *English for Specific Purposes* 9 (2), 109–21.

Scardamalia, M. & Bereiter, C. 1987. Knowledge telling and knowledge transforming in written composition. In: Rosenberg, S. (Ed.) *Advances in Applied Psycholinguistics, Vol. 2: Reading, writing, and language learning.* Cambridge University Press, 142–175.

Shaw, S. D. & Weir, C. J. 2007. *Examining Writing. Research and practice in assessing second language writing.* Cambridge: University Press.

Sigott, G., Gassner, O., Mewald, C. & Siller, K. 2007. *E8 Standardstests. Entwicklung der Tests für die rezeptiven Fertigkeiten: Überblick. LTC Technical Report 1.* Language Testing Centre, Alpen-Adria-Universität Klagenfurt. Available as download from http://www.bifie.at/publist-07-02-19 [24 June, 2011]

Tankó, G. 2005. *Into Europe. The Writing Handbook.* Budapest: Teleki László Foundation.

Verordnung der Bundesministerin für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen. BGBl. II Nr. 1/2009 v. 2.1.2009. Available as download from http://www.bmukk.gv.at/schulen/recht/erk/vo_bildungsstandards.xml [14. April, 2011]

Weigle, S. 1994. Effects of Training on Raters of ESL Compositions. In: *Language Testing* 11 (1), 41–69.

Weigle, S. C. 2002. *Assessing Writing.* Cambridge: University Press.

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

Appendix

Prompt Interpretation: Long Prompt

You have just moved to another town/village. Write a **letter** to your American/English friend in which you tell him/her about your new situation.

Inform him/her about
- your new place of living. (1)
- the reason for moving. (2)

Describe
- the town/village you're living in now (buildings, people,…). (3)
- your new home. (4)

Tell him/her about
- the first days of your 'new life' (new school, teachers,…). (5)
- how you feel about your new situation. (6)

Task Achievement

Content

- Although the first content point can be dealt with very briefly (new place), the second point (reasons for moving) allows for some elaboration.
- The next three points (describe town/village, new home, and first days) definitely allow for elaboration and we expect somewhat detailed descriptions (points 3, 4, 5).
- The last point about feelings is more difficult because this requires some creative thinking, which is why we do not expect much elaboration. If there is elaboration here, it needs to be specially recognised.

Degrees of Task Achievement

In this long text, all content points need to be mentioned for **complete task achievement** and three elaborated. Elaboration in the last point about feelings cannot be expected as this is considered to be difficult to achieve at the level of the test takers.

Good task achievement can be awarded even if one content point is not dealt with. '85%' means that five content points must be mentioned. Depending on the quality of elaboration, we would expect at least two points to be elaborated for good task achievement.
A second option for band 5 is to have all content points mentioned and one or two elaborated.

For **sufficient task achievement** we would expect four content points to be mentioned and one or two to be elaborated, depending on the quality of elaboration.

A second option for band 3 is to have all content points mentioned, but no elaboration.

Some task achievement would be given if the message was clear enough to convey the information that the person has moved to another place and some brief information about two other content points to reach the 50%.

Text Type Requirements

- The text type is clearly marked as a personal letter, which is why we expect an informal opening and closing formula like ***Dear Jim/Hi Sue*** and ***Yours/Love*** etc.
- The register needs to be ***informal.***
- Text type requirements are **not met** if there is no opening and/or closing formula AND if the register is not appropriate (e.g. rude language or mismatch of informal situation and formal language). This will lead to downgrading.
- Missing or incorrect opening and/or closing formula will lead to downgrading by one band.
- Inappropriate register/tone alone will lead to downgrading by one or two bands depending on the quality of the text. The same applies to an impolite or offensive tone.

Coherence and Cohesion

In a long text we definitely expect paragraphing. Since the content points are grouped into three units, the test takers are provided with sufficient hints as to how the text could be structured into paragraphs. A lack of paragraph organisation in bands 3 to 7 leads to downgrading by one band.

- Test takers need to use paragraphs and some cohesive devices to reach band 7. The text must be clear and coherent at sentence, paragraph and text level.
- For band 5, cohesion should be achieved at least at sentence level, and longer stretches should show some paragraph level coherence and cohesion. For example, this could be demonstrated in the description of the new place and/or home.
- For band 3, we would at least expect the test takers to use simple connectors (e.g. *because, but, and, then)* when giving reasons for moving or for the description of places and/or the first days at school. So we can expect sentence level cohesion, but problems at paragraph level. The text might not be fully coherent.
- For band 1, it is enough if the sentences are linked with very basic connectors. We do not expect paragraphing or textual coherence for band 1.
- If the letter is written in chunks rather than sentences, we suggest that there is not enough assessable language for coherence and cohesion.

Grammar

In this long text, writers are expected to inform about the move to another place making use of past verb forms. The reason for moving can either be linked with this information and also realised in the past, or connected to the information about the new place and therefore feature present tense verb forms. The description of the new place, the new home and the new life will most probably elicit present tense verb forms. Very good solutions might include comparisons between the new and the old place, the new and the old situation. However, these cannot be considered compulsory because they are not explicitly required by the task.

Vocabulary

The instructions and content points should entice the writer to use a lot of familiar content words to describe the new situation. We would expect the writer to mention the name of a country, city, town or village and use content words to describe it, the buildings, people, neighbours, and possibly the countryside there. The writer has the opportunity to describe his/her home and words relating to types of houses, household rooms, furniture, and perhaps a garden, will most likely be in the text. We would certainly expect the writer to use a range of adjectives such as qualifiers

to describe appearance, condition, and/or shape; also quantifiers to describe size, in order to compare the present situation to the former situation. The writer should give a reason for moving and he/she will probably use vocabulary relating to a past event in the family: a parent getting/losing a job, starting a new life, moving back to a former place of residence, or a change in the family structure due to a divorce.

The writer should also write about his/her new life and we can expect some vocabulary about school, teachers, classmates, subjects, classroom, and/or homework to appear in the text. It may well be that the writer chooses to ignore writing about school and interprets "the first days of your 'new life'" by referring to out of school activities such as buying/moving furniture, meeting and playing with new friends, exploring the new area and other free-time activities. Once again, we would expect the sentences to contain a mixture of qualifiers and quantifiers.

Finally, we would expect some of the sentences about the new situation to include adjectives to describe the positive, ambivalent and/or negative feelings he/she is experiencing in the new situation.

Prompt Interpretation: Short Prompt

Your friend's birthday party was a few days ago. Write an **email** to tell him/her that you liked the party.

- **Tell** him/her why you liked the party. (1)
- **Tell** him/her what you liked best. (2)
- **Ask** your friend when you are going to meet again. (3)
- **Suggest** something for the next weekend. (4)

Task Achievement

Content

- Point 1: The writers need to give one or several reasons for liking the party. It is not enough to say that they liked the party but this should include giving reasons.
- Point 2: This can be done in a more or less elaborate way depending on the writer's choices and abilities.
- Point 3: This is simple enough and will be dealt with very briefly. Most probably it will immediately lead on to point 4.
- Point 4 is expected to be the most challenging one as it requires some thinking and a minimal amount of creativity. Writers need to include at least one suggestion referring to the following weekend.

Degrees of Task Achievement

With short tasks, the relevant descriptor needs to be interpreted very strictly. In fact, with four content points required it makes little sense to tolerate a missing one. It should be taken for granted that ALL four content points must be dealt with for complete or good task achievement. This would normally mean a text length of around 60 words.

- It will be the degree of elaboration that distinguishes between complete and good task achievement.

- With one content point missing, good task achievement is no longer possible, but band 4 is.
- If one content point is missing, we can award band 3, i. e. sufficient task achievement (just over 70%). If there is some elaboration, it is band 4.
- If two content points are missing, we cannot expect much elaboration either and it seems to point towards band 1.
- If, however, there is good elaboration of one content point, it is band 3. If elaboration is weak, it is band 2.

Text Type Requirements

- We can expect an informal opening and closing formula like ***Dear Jim/Hi Sue*** and ***Yours/Love*** etc.
- For band 7 there must be an appropriate opening and closing formula and the register must be informal and demonstrate the familiarity indicated by *your friend.*
- Problems with the beginning or ending of the text or minor problems with the register would lead to downgrading by one band.
- When the problems become more obvious, task achievement will drop by two bands.
- Text type requirements are not met when there is no opening and/or closing formula AND when the register is not appropriate. (Downgrade by two bands if the problems with register are very serious).

Coherence and Cohesion

- We do not expect any paragraphing with short tasks although good writers will use paragraphing to structure their short texts as well. So some credit should be given for successful paragraphing.
- Demonstrating this dimension in short texts is rather difficult, but we can expect the writers to deal with the four content points in the order given. Implicit and explicit linking (connectors) is not easy to place in this kind of text and should be specially acknowledged when it is used successfully. Content points 1 and 2 suggest the use of *because* and/or *and.*

Grammar

This short task asks for feedback about a past event. The description of what was liked and what was liked best opens up possibilities for comparisons or even the superlative. However, any other solution that implies the information about the liked and best liked activity at the party needs to be acknowledged. Moreover, the task requires a question for a meeting in the future and a suggestion for an activity for the next weekend. While the question is very likely to elicit a future verb form (most probably the *going to* future as given in the prompt), the suggestion can also be realised in the present or making use of modal verb forms. The latter could also come across as another question followed by a suggestion or a statement *(What about next weekend? My brother has a birthday party. Would you like to come?)*

Vocabulary

We expect the writer to use adjectives such as *great, cool, funny* and/or to refer to nouns such as *the games, music, food, lights, friends that were at the party* in the response to the first content point. The writer will probably rephrase the second content point and add similar vocabulary to state clearly why he/she liked the party best. The response to the third content point will most likely contain one of the following: *When can I see you again? When are we going to meet again? What are you doing at the*

weekend? The final part of the text will most likely be a combination of a suggestion and a question containing words related to common teenage leisure activities such as *meeting friends, going shopping/to a party,* and/or *playing/watching football* etc. As the last content point refers to the weekend, we can expect the writer to include words such as *weekend, Friday, Saturday, Sunday, morning, afternoon, evening* and/or *night.*

Inventory of Functions, Notions and Communicative Tasks[1]

Conveying Emotions, Feelings

- criticising and complaining
- expressing needs and wants
- expressing preferences, likes and dislikes
- making and responding to apologies and excuses
- paying compliments
- sympathising
- talking about feelings

Informing, Asking

- asking and answering questions about personal possessions
- asking and telling people the time, day and/or date
- asking for and giving information about routines and habits
- asking for and giving personal details: (full) name, age, address, names of relatives and friends, occupation, etc.
- asking for and giving simple information about places
- completing forms giving personal details
- expressing (in)ability in the present and in the past
- making predictions
- talking about (im)probability and (im)possibility
- talking about food
- talking about future or imaginary situations
- talking about future plans or intentions
- talking about one's health
- talking about the weather
- talking about past events and states in the past, recent activities and completed actions
- talking about what people are doing at the moment
- writing diaries giving information about everyday activities

Convincing, Persuading, Expressing Opinions

- asking and giving/refusing permission to do something
- drawing simple conclusions and making recommendations
- expressing agreement and disagreement, and contradicting people
- expressing degrees of certainty and doubt
- expressing obligation and lack of obligation
- expressing opinions and making choices
- expressing purpose, cause and result, and giving reasons
- giving advice
- giving warnings and prohibitions
- persuading and asking/telling people to do something

1 Adapted from *PET Handbook* 2009 and *KET Handbook* 2009. The list is not exhaustive, but serves to illustrate aspects of the Construct Space.

Entertaining, Pleasing

- talking about past events and states in the past, recent activities and completed actions
- talking about what people are doing at the moment
- understanding and producing simple narratives

Keeping in Touch

- expressing and responding to thanks
- giving and responding to invitations
- making and granting/refusing simple requests
- making and responding to offers and suggestions
- writing letters/emails giving information about everyday activities
- writing letters/emails giving personal details

Describing

- buying and selling things (costs, measurements and amounts)
- describing education and skills
- describing people (personal appearance, qualities)
- describing simple processes
- identifying and describing accommodation (houses, flats, rooms, furniture, etc.)
- making comparisons and expressing degrees of difference

Giving Directions and Instructions

- asking for and giving travel information
- asking the way and giving directions
- following and giving simple instructions
- identifying and describing simple objects (shape, size, weight, colour, purpose or use, etc.)
- talking about how to operate things