R&E-SOURCE http://journal.ph-noe.ac.at
Open Online Journal for Research and Education
Ausgabe 1, April 2014, ISSN: 2313-1640

Practice what you preach – test what you teach

Testing English as a foreign language in heterogeneous groups

Claudia Mewald[*]

Abstract

This paper discusses criterion oriented assessment in the context of competence oriented foreign language education and standardised testing. It provides information about test design, scale development and the application of a 4.0 scale system in heterogeneous settings. Moreover, the role of formative and summative assessment in planning teaching and learning with the end in mind is explained.

Was wir testen und was wir unterrichten

Schularbeiten im Fach Englisch in der Neuen Mittelschule

Zusammenfassung

Dieser Artikel bespricht kriterienorientierte Beurteilung im Kontext des kompetenzorientierten Fremdsprachenunterrichts und des standardisierten Testens. Er informiert über die Entwicklung von Schularbeiten und von Bewertungsskalen sowie über die Anwendung eines 4.0 Systems in heterogenen Klassen. Außerdem wird die Rolle der formativen und summativen Beurteilung im Rückwärtigen Lerndesign[1] geklärt.

Keywords:	*Schlüsselwörter:*
Criterion oriented assessment	Kriterienorientierte Beurteilung
Tests in heterogeneous groups	Schularbeiten Gruppen
Testing and assessment	Testen und Beurteilen

1 Introduction

The educational system in Austria in general and the subject English as a foreign language (EFL) in particular have recently been facing substantial change and innovation in various aspects. Standards based on the Common European Framework of Reference for Languages (CEFR) were implemented, followed by the development of standardised tests. However, discussion arose about the role and value of testing and assessment in the context of EFL in heterogeneous learner groups at Austrian New Middle Schools (NMS). This rapid change process has left many teachers in limbo, overwhelmed by the many claims and tasks that appear to be impossible to solve because of their seemingly controversial nature.

However, what may seem to be controversial at first sight is actually an approach that aims at the same global goal, which is to change EFL learning and teaching in a way that all learners are provided with the opportunity to learn to use the language communicatively and effectively in real life.

Since the official enactment of the Austrian "Bildungsstandards" in 2009, competence oriented education has made its way into many classrooms. For the subject English "standards" were provided in the form of can-do statements. They were adapted from the CEFR (Council of Europe, 2001) and relied on a competence model (BIFIE, 2011-2014a; BMBF, 2009) in line with a communicative approach to foreign language learning and teaching. With communicative competence having been the commonly acknowledged aim of EFL since the

[*] Pädagogische Hochschule Niederösterreich, Mühlgasse 67, 2500 Baden.
E-Mail: claudia.mewald@ph-noe.ac.at

onset of Communicative Language Teaching (Hymes, 1972), it seemed that neither teachers nor material developers would have to make substantial changes to what they had already been doing, using or publishing.

Together with the development and piloting of the standards for English (E8 standards), the curriculum for foreign languages was revised accordingly (BMBF, 2008). The list of functions and notions (grammatical and lexical elements) was replaced by descriptions of competences in listening, reading, writing, oral production, and spoken interaction at CEFR levels ranging from A1 to B1. Moreover, a standardised testing instrument was developed that should provide opportunities for measuring the competences in EFL reached at the end of year eight with the goal of system monitoring to guide and plan quality development and assurance in education (BIFIE, 2011 - 2014b). The last component to be added to the change process was differentiation according to the learners' readiness levels, interests, and preferred modes of learning (Rock et al., 2008; Tomlinson, 1999).

This paper aims to shed light on the commonalities and differences within the various approaches, demands, and perceptions. It also suggests a modus vivendi that might serve in the search of a beneficial approach to testing and assessment in heterogeneous learner groups.

2 Testing and assessment – an impossible match?

Although the terms "testing" and "assessment" are often used interchangeably or even as a lexical chunk "as if they were a single entity" (Allan, 1999, p. 4), they also have dichotomous connotations. While language tests generally hold the notion of formality, validity as well as reliability, they also tend to raise associations of unease or even anxiety, frequently perceived with mistrust in their trustworthiness as regards their outcomes by lay people. Too many pitfalls may have influenced the results: the test might be testing test wiseness rather than real language competence, the test taker may have had a bad day, or test fairness might be doubted etc. Nonetheless it has to be acknowledged that professional language testing has a long tradition as well as ample expertise, and that it is one of the best researched fields in education.

While testing is generally considered to deliver summative feedback, assessment is most commonly associated with a soft and less threatening mode of collecting data for the purpose of providing formative feedback. In this sense it is clearly different from testing. If assessment is used as a formative tool, it provides information that is used *for* learning rather than *about* learning (outcomes), and it focuses on the process with the best possible product in mind. However, formative assessment may also make use of testing when it comes to diagnosing strengths or weaknesses in order to plan learning or teaching. Such "tests" do not contribute to scoring, but they have diagnostic power through the feedback they can provide. For any EFL test to hold the potential of useful feedback or diagnosis, it has to be based on a theory of language that is reflected in its construct. If learning and teaching aim at competence with successful performance in real life, the planning process should focus on global goals which are defined together with the expected outcomes and their assessment in a backward design prior to instruction. This way, any testing or assessment will use criteria to evaluate performances and thus provide the opportunities for feedback that can be used for planning learning and teaching.

3 Criterion oriented testing

Criterion oriented[2] testing "examines the level of knowledge of, or performance on, a specific domain of target behaviours (ie the criterion) which the candidate is required to have mastered."[3] (Davies, et al., 1999, p. 38)

In EFL, the curriculum (BMBF, 2008) and the E8 standards (BMBF, 2009) provide the performance descriptors, which are expected to be mastered at the end of year eight of lower secondary education. While the levels for years five to six are to progress from A1 to B1 according to curricular guidelines, the specific criteria for the five skills are not explicitly stated. However, the close connection of the curricular guidelines and the E8 standards to the CEFR (Council of Europe, 2001) offers the opportunity to exploit the CEFR descriptors "for the development of criterion-referenced assessment" according to "local systems" and "elaborated by local experience" (Council of Europe, 2001, p. 30). This provides the basis for the teachers' work when developing tests for formative or summative purposes.[4]

4 Teaching and testing in heterogeneous groups

An analysis of Austrian schoolbooks suggests that the most recently published teaching materials have already been adapted to the new curricular guidelines, the E8 standards and the requirements of NMS, while the testing materials that come with the courses have not yet been changed so far. Although they offer tasks and items aiming at three levels of difficulty that go back to the system of streaming in general secondary education, their diagnostic potential is widely unused because the materials do not yet provide the information needed for the results to be diagnostic. In order to do so, the items would have to be identified to test certain abilities defined in a construct that is based on a theory of language.

The theory that underlies most teaching and learning in EFL is that of communicative competence. As suggested in the CEFR, "communicative competence … has the following components:

- linguistic competences;
- sociolinguistic competences;
- pragmatic competences." (Council of Europe, 2001, p. 108)

All of these competences should come into play when teachers design the learning and teaching of EFL and they should also be considered in the learners' assessment in order to "practice what we preach"[5] and to test what we teach. If testing and/or assessment refer to linguistic competences alone (which is often the case), the construct is incomplete, and the feedback we are able to give can only be equally fragmentary. Moreover, if the only feedback a test result can provide is the number of points or a percentage scored, the concept of feedback *for learning* is not fulfilled.

5 From teaching and learning to testing

As suggested by Schlichterle & Weiskopf-Prantner (2013, p. 3), it is essential to design authentic tasks to be able to test competences that have been developed in target oriented teaching and learning processes. One may say that authenticity in testing is not exactly an easy goal to be achieved. Where and how would anybody face a testing situation in real life situations outside classroom walls? However, if authenticity refers to the tasks, then it relates to the interaction in the classroom normally carried out by learners and teachers, which is the real life scenario in an educational context. Thus, if the tasks engage the learners in authentic situations, the output they produce can be considered authentic as well. Taking this into consideration, we are testing in an authentic way, when we test what we teach and what we teach is what we would normally do in real life.

If one follows the eight features of competence oriented foreign language teaching, then teaching and learning should have a clear focus on successful functioning in society as well as on life skills; that they should have a task- or performance-centred orientation and be implemented through modularised instruction with the outcomes made explicit a priori. Continuous and ongoing assessment would consequently be a tool to demonstrate the mastery of performance objectives pursued in individualized and student-centred educational contexts (Richards & Rogers, 2001).

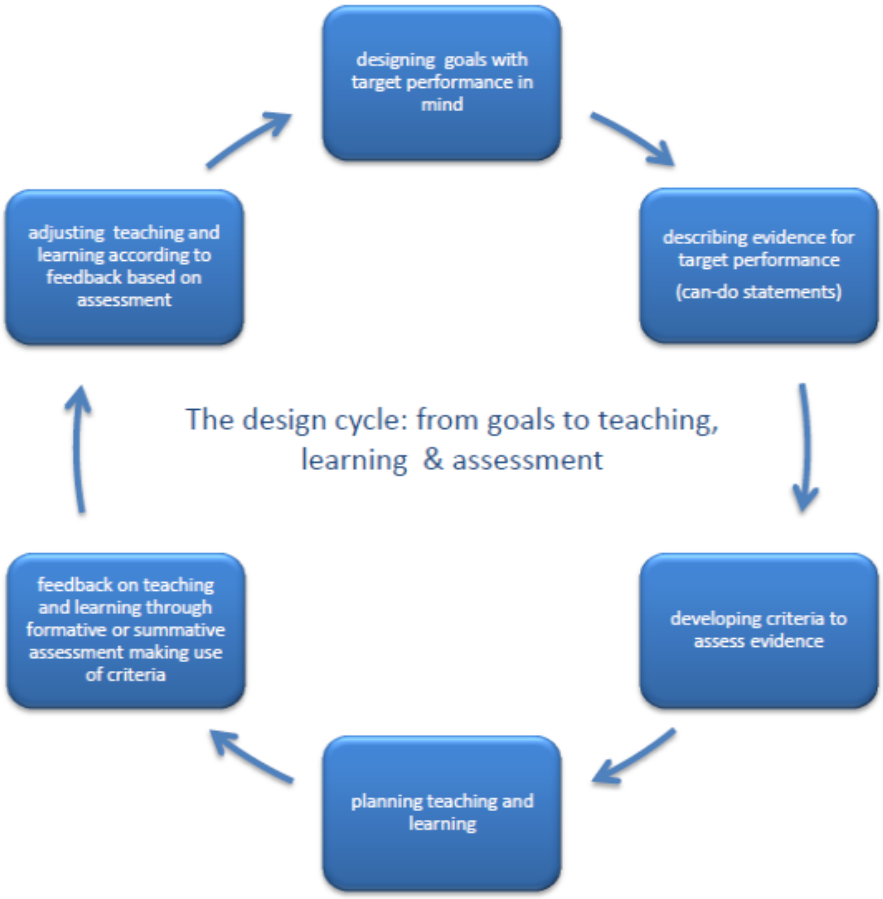

Fig. 1: The design cycle: from goals to teaching, learning and assessment

In order to assess individualised and student-centred performances, criterion oriented assessment seems appropriate because it allows judging the performances according to set criteria that can be directly translated into valuable feedback at defined levels.

6 Creating assessment scales

Criterion oriented tests report on test takers' abilities in relation to criteria in assessment scales, also called rating scales or rubrics. This way they give information about what test takers can or cannot do. The can-do statements used in scales should be based on a theory of language or a specific language skill, such as reading or listening etc. Can-do statements provide descriptive results rather than numerical scores. This way, successful performances can be described with reference to linguistic and non-linguistic criteria. For example, speaking performances in E8 standards testing are assessed according to linguistic-criteria, which include fluency, coherence and cohesion, lexical and grammatical range and accuracy. Non-linguistic criteria include the achievement of the task through assessing thematic development, propositional precision and turn-taking skills (Mewald et al., 2013). The different criteria are graded in order to give feedback about the level of fulfilment. While E8 testing relies on scales with seven levels in writing and speaking, NMS practice advises one to make use of rating scales particularly developed for each design and to follow a 4-Point Scale system like Marzano's (Marzano, 2006; Marzano, 2010).

Assessment scales can be used before teaching to diagnose prior knowledge, which is an important aspect of backward design (Wiggins & McTighe, 2005), during teaching to plan next steps in teaching and learning (formative assessment), and for grading (summative assessment) as soon as teaching and learning have been completed.

To be effective, Marzano suggests the descriptors in the scales or rubrics should reflect learning progressions in "increasingly more sophisticated levels of knowledge or skill." (Marzano, 2010, p. 42)
Thus a generic form of a scale would look like this:

Score 4.0	More complex goal
Score 3.0	**Target learning goal**
Score 2.0	Simpler goal
Score 1.0	With help, partial success at score 2.0 **and** 3.0
Score 0.0	Even with help, no success

Table 1: Scale or rubric adapted from Marzano **(2006, p. 45)**, emphasis added

Learners, who reach the target goal, i.e. if they demonstrate competence regarding the descriptor at band 3.0. Performances, which indicate that the learning process has gone beyond the target goal and reached a more sophisticated and more complex level, result in a score 4.0. If a performance does not yet reach the level of the target goal but the one of the simpler goal, its assessment will be 2.0. If help was needed to reach partial success in 2.0 or 3.0 goals, the performance will be awarded 1.0. Performances that are considered unsuccessful despite the support that has been provided are assessed 0.0.

Taking into consideration that rubrics reflect learning progressions, it seems meaningful to design goals that reflect this progression and can guide teaching and learning. This way, the learners also know how far they have already progressed in the process at any time.

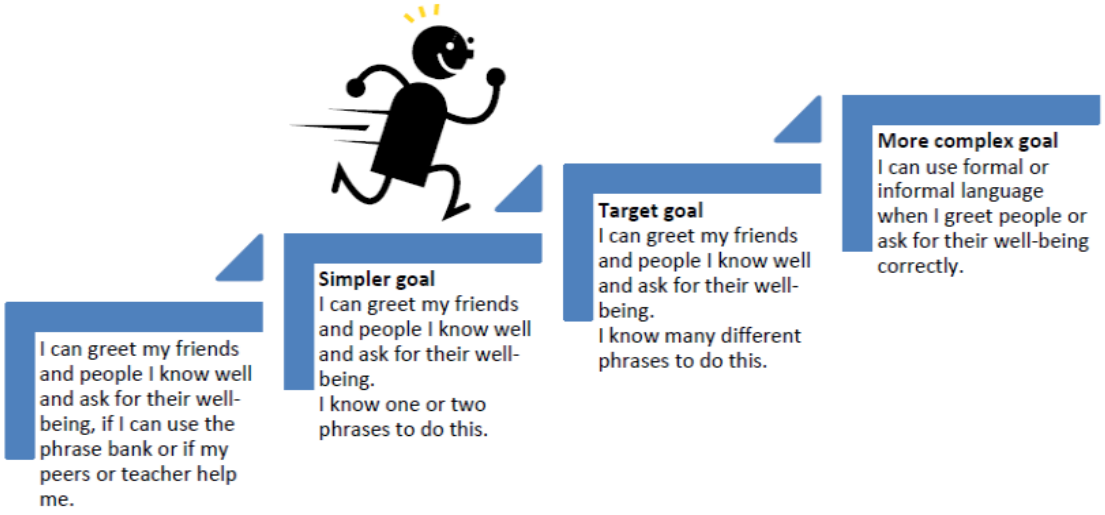

Fig. 2: Goals that reflect learning progression

Goals that reflect learning progressions not only provide the opportunity to guide teaching and learning, but also to be translated directly into criteria for self- or peer-assessment. Moreover, they should be the basis for the development of criterion oriented assessment scales for summative purposes. This way teachers will practise what they preach, and they will test what they teach.

When designing scales at four levels, it is important to start from the description of the expected target performance and the essential indicators, which would describe the criteria for mastery.

Next, the level below mastery is described in a way that the criterion would enable the assessors to differentiate between performances that reach mastery level and ones which are unable to do so yet. To avoid a pass-fail scenario, which would most certainly not be perceived positively and yet again lead to anxiety or stress, the criteria below mastery, i.e. those at levels 2.0 and 1.0, should be clearly signposted as stepping-stones towards mastery rather than failure.

Effective scales therefore use level descriptors that clearly discriminate between mastery and non-mastery, but at the same time, all descriptors must be phrased as positive can-do statements. This is not to camouflage problem areas, but to provide the opportunity to make explicit assessments or to help understand what has already been mastered and what is still to be achieved. Thus, if diagnostic feedback should be given, the descriptors must also be designed in a way that they provide useful information about the stage in the process that a certain performance can be related to. Scales designed in this way assume that learners who can fulfil tasks at a certain level will also have mastered the lower levels.

	The learner / test taker can….	In the context of EFL education at lower secondary schools the input texts would range from CEFR levels A1 to B1. This offers opportunities for progression in complexity and difficulty.
4.0	infer information that is not explicitly stated in a text.	
3.0	**identify essential information that is explicitly stated in a text and distinguish it from supporting detail.**	
2.0	identify concrete information that is explicitly stated in a text.	
1.0	identify the main idea (gist) of a text (e.g. its explicit purpose, the text type, the topic …)	

Table 2: Example for an assessment scale for receptive language performances

It can be seen that the 1.0 descriptor in Table 2 does not refer to "help" as suggested in the generic scale form (see Table 1). In a formal testing situation providing help would raise issues of test administration. Learners who are given help would have to be separated from their peers to prevent access to the help by test takers, who are not expected to make use of it. To avoid this situation, items at four levels of complexity and/or difficulty seem more appropriate in formal testing.

7 Designing EFL tests

In heterogeneous settings, criterion oriented assessment will only be effective if EFL tests contain items at different levels of complexity and/or difficulty. The easy approach to differentiate items by quantity or difficulty alone should not be considered because this would not reflect a competence oriented approach of assessing language performances. However, the caution as in regard to quantity does not relate to the total number of items in a test, however. Tests with a higher number of discrete items provide more reliable results than ones with just a small number of items. Moreover, test takers get more opportunities to show mastery in tests with more items.

"In tests of speaking and writing, it is possible to argue that the rating process itself can be criterion-referenced, since most descriptors for rating criteria contain, essentially, definitions of adequacy, and raters have to judge whether or not the candidate meets the standard for that criterion. Although this is itself not straightforward, it is much more complicated to apply such a principle to discrete-point tests of grammar or vocabulary, or even to integrative cloze tests intended to measure reading ability. What is an adequate score on this grammar test? What can be considered to be a pass on this cloze test?" (Alderson et al., 2004, p. 157)

In the context of competence oriented foreign language education, criterion oriented assessment thus raises an issue that is crucial to test design. According to Richards & Rodgers competency based assessment only addresses specific language skills learnt during a course, and the criteria that form the basis for the assessment only refer to the "essential skills, knowledge, attitudes, and behaviours required for effective performance of a real-world task or activity" (Richards & Rogers, 2001, p. 144). If this is the case, then discrete-item testing of decontextualized vocabulary or grammar items cannot be part of competency based tests. In fact, grammar and vocabulary are components of any productive writing or speaking performance, which is why there does not seem to be an urgent need to isolate them in testing.

However, if validity issues are taken seriously, discrete-item testing still seems to be an appropriate option to get reliable and valid information about reading or listening skills because once writing comes into play it can no longer be securely said that it was that the reading or listening that failed rather than the production of the response. In test design we should therefore consider a substantial number of closed items in reading or listening tasks, while the writing tasks would have to require text production rather than gap-filling or completion.

If scores generated through criterion oriented assessment scales have to be converted into grades, the following scheme should be followed as suggested by Schlichterle & Weiskopf-Prantner:

Scores	Grades
At least half of the scores are 4.0, the rest are 3.0	Sehr gut
3/4 of the scores are 3.0 or 4.0, the rest are no lower than 2.0	Gut
At least 40% of the scores are 3.0 or 4.0 and the remaining 60% are no	Befriedigend

lower than 2.0	
At least half of the scores are 2.0 or above	Genügend
At least 1/4 of the scores are 2.0 or above and the rest are no lower than 1.0	Befriedigend in grundlegender Allgemeinbildung
At least 3/4 of the scores are 1.0 or 1.5 and the rest are no lower than 0.5	Genügend in grundlegender Allgemeinbildung

Table 3: Entscheidungsgrundlage für die Ermittlung einer Gesamtnote für die 7. & 8. Schulstufe **(Schlichterle & Weiskopf-Prantner, 2013, p. 43)**, translation

While this scheme seems perfectly appropriate for a final grade based on many scores collected throughout a school year or term, it seems a bit too strict for a single test. Firstly, test takers have to fulfil all items in a single test, which will contain items from 1.0 to 4.0. Secondly, it may happen that test takers make some mistakes in items at 1.0 to 2.0 level, but they may still get all or most of the 3.0 and 4.0 items right. Therefore, the following adaptation is suggested for tests:

Test takers are awarded ……	**if they score …….**	**and if they score ….**
Sehr gut	at least 50% of the 4.0 items	and 90% of the remaining 3.0, 2.0 and 1.0 items
Gut	at least 75% of the 3.0 or 4.0 items	and 90% of the remaining 2.0 and 1.0 items
Befriedigend	at least 40% of the 3.0 or 4.0 items	and 90% of the remaining 2.0 and 1.0 items
Genügend	at least 50% of the 2.0 items or above	and 90% of the remaining 1.0 items
Befriedigend in grundlegender Allgemeinbildung[6]	at least 25% of the 2.0 items or above	and 90% of the remaining 1.0 items
Genügend in grundlegender Allgemeinbildung	at least 75% of the 1.0 items	

Table 4: Converting scores into grades in tests

Any EFL test should have a reading, a listening and a writing component and all three skills should be tested and weighted equally. Therefore, test designers have to make sure that the total number of scores generated from reading, listening, and writing tasks is the same. Equally, the total number of scores within each of the four categories from 4.0 to 1.0 must not vary. This suggests that the number of scores in any category is 4 or a multiple of 4.

Fig. 3: Number of scores in tests

If a writing scale with four dimensions and four levels is used, it will produce a maximum of 16 scores (see 4.1.1). Therefore, it would make sense to design reading and listening tasks with the same number of items, i.e. 16 each. This would add up to 48 scores in total with 12 possible scores in each band. The number of reading or listening tasks can be varied, but the total number of items within each skill and band should be consistent.

For example, a test could have two reading tasks but four listening tasks as suggested in the table below.

Reading 1	Reading 2	Listening 1	Listening 2	Listening 3	Listening 4
Item 1: 1.0	Item 9: 1.0	Item 1: 1.0	Item 5: 2.0	Item 10: 1.0	Item 13: 3.0
Item 2: 1.0	Item 10: 1.0	Item 2: 1.0	Item 6: 3.0	Item 11: 1.0	Item 14: 2.0
Item 3: 2.0	Item 11: 2.0	Item 3: 2.0	Item 7: 4.0	Item 12: 2.0	Item 15: 3.0
Item 4: 2.0	Item 12: 2.0	Item 4: 4.0	Item 8: 4.0		Item 16: 4.0
Item 5: 3.0	Item 13: 3.0		Item 9: 3.0		
Item 6: 3.0	Item 14: 4.0				
Item 7: 3.0	Item 15: 4.0				
Item 8: 4.0	Item 16: 4.0				

Table 5: Exemplified distribution of items in a test

When marking tests, teachers will count scores or ticks in each band from 1.0 to 4.0. Applying the conversion table (see Table 4) the following results will be produced, if the total score in a test is 48:

In a test with a total score of 48 the test takers are awarded …	**if they score …**	**and if they score …**
Sehr gut	at least 6 of the 4.0 items	and at least 32 of the remaining 3.0, 2.0 and 1.0 items
Gut	at least 18 of the 3.0 or 4.0 items	and at least 21 of the remaining 2.0 and 1.0 items
Befriedigend	at least 9 of the 3.0 or 4.0 items	and at least 21 of the remaining 2.0 and 1.0 items
Genügend	at least 6 of the 2.0 items or above	and at least 10 of the remaining 1.0 items
Befriedigend in grundlegender Allgemeinbildung[7]	at least 3 of the 2.0 items or above	and at least 10 of the remaining 1.0 items
Genügend in grundlegender Allgemeinbildung	at least 9 of the 1.0 items	

Table 6: Converting scores into grades in tests with a total score of 48 items and 12 items in each band

While discrete items can be ticked off and counted if correct, scores from writing scales have to be treated differently.

8 Assessing written performances in tests

In real life, successful communication is the most important criterion in the assessment of productive skills. However, it is not always only important to get a message across and to fulfil a task. There are situations when accuracy is crucial or when range decides upon success or failure.

Therefore, writing assessment scales have various categories in addition to task achievement. In E8 Testing and in IKM Writing (Informelle Kompetenzmessung), the writing rating scales feature the following dimensions: Task achievement, Coherence and Cohesion, Grammar, and Vocabulary[8]. While E8 Testing uses seven bands, IKM works with four. Since both scales offer descriptors for four bands, they can be applied in the 4.0 system used in NMS. Using a rating scale reliably requires careful training, especially in the dimension of task achievement. Therefore, practical use in teaching and testing situations suggested the development of a less elaborate scale. The below 4.0 Writing Assessment Scale has proven to be less difficult in use, especially for teachers who are less experienced in the use of criterion oriented rating scales.

Task Achievement How well does the writer convey the message? Are there any details? Does the writer stick to the text features?	- message/content meaningful and completely successful - several relevant and interesting details - layout and format fully appropriate, text features met	4.0
	- message/content meaningful and mostly successful - some details - layout, format, text features support text	3.0
	- message/content not always meaningful/clear; reader left with questions - hardly any details - layout, format do not support text; text features hardly met; limited length	2.0
	- message/content hardly meaningful/clear - no details - layout, format not appropriate; text features not met; very limited length	1.0
Coherence and cohesion	- text well organised with strong beginning – middle - end	4.0

How well is the text organised? Are beginning, middle and end clearly and effectively marked? Does the text flow, i.e. is it coherent on sentence and paragraph level? Are paragraphs coherent and visibly marked?	- sentence starters, linking words and good sentence level coherence - paragraphs largely coherent	
	- text organised with sense of beginning – middle - end - sentence starters, linking words - ideas clustered but paragraphs not visibly marked	3.0
	- text only loosely organised - no transitions - very little sentence level cohesion - ideas not linked	2.0
	- text not organised and confusing - no sentence level cohesion	1.0
Linguistic range How varied are lexical elements and grammatical structures? How well do they support the message and contribute to the success of the text?	- linguistic range helps to convey message effectively - consistent variation of compound and complex sentences - choice of words, phrases and structures make the text interesting and engaging	4.0
	- linguistic range is appropriate and purposeful - simple compound and complex sentences	3.0
	- linguistic range sometimes limited; repetitions make the text flat and vague - sentence structures show little variation and are repetitive	2.0
	- limited linguistic range makes the text obscure - no variation in sentence structure	1.0
Accuracy How correct is the language? Do mistakes (lexis or grammar) impair understanding?	- language mostly correct - few mistakes usually without impact on understanding	4.0
	- language sometimes incorrect - some mistakes mostly without impact on understanding	3.0
	- language often incorrect - several mistakes with some impact on understanding	2.0
	- language mostly incorrect - many mistakes with strong impact on understanding	1.0

Table 6: 4.0 Writing Assessment Scale

No matter what rating scale is applied, if it is based in a 4.0 system it can be used with the conversion rules shown in Tables 4 or 6.

Written performances will meet certain criteria described in the scales and the scores will be awarded accordingly. For example, if a performance is awarded band 4.0 in task achievement, ticks will also be given in 3.0, 2.0 and 1.0. This reflects the notion that any written performance at band 4.0 has also met the criteria for 3.0, 2.0 and 1.0. Accordingly, if a performance is assessed 2.0, band 1.0 will also be ticked.

This way a writing performance can reach a maximum of 16 ticks when assessed with a 4-dimensnional assessment scale based on 4 bands.

9 Teaching with the end in mind and testing what has been achieved

Teaching, with the end in mind, means one must be clear about the objectives of teaching and learning. Therefore, backward design and competence oriented foreign language education make a perfect match. In real life situations, we are always clear about the expected outcome. If we take a car to a garage because it does not work, Wormeli suggests, we know exactly what we expect the mechanic to do: “There’s something wrong with this car. If you can figure out what it is and fix it, I’ll pay you.” (Wormeli, 2006, p. 21)

The mechanic will tell us what kind of tests will be necessary to find out about the problem, and once it is detected, an estimate will be provided to tell the customer what parts are needed and how much the repair work will cost.

In competence oriented teaching, we face a similar situation. We know what the learners should be able to do at the end of a phase of teaching and learning, and we should not let them in the dark about our goals. Like the customer in the garage, we would also like to know what materials and how much time might be needed to achieve a goal. The only way to stipulate a realistic estimate is to devise a diagnostic test. Knowing the prior knowledge of our learners as well as their gaps makes a better basis for planning than a vague estimate of how easy or difficult a task might be for a particular group of learners. What may have gone well and fast with one cohort of learners might take ages with another. If we know the reason why, ideally before planning teaching, and testing, our interventions can be more targeted and the expected evidence for and level of mastery will also be more realistic. This would be the formative aspect of assessment.

Looking at summative assessment and competence oriented foreign language education in an era of standardisation, one may think that the idea of planning learning and teaching according to the learners' needs and their levels of readiness could be severely challenged, especially in heterogeneous groups.

This need not be the case. "In a differentiated classroom assessment guides practice" (Wormeli, 2006, p. 20) and if we practise what we preach and test what we teach, testing will lose the threat of the unexpected and new and become a means of celebrating success instead. Wormeli has a radical suggestion: we should show the learners the end-of-unit test at the beginning of a teaching unit and clarify each part of it. This will make the learners more targeted and attentive because they achieve more if they have a clear picture of the expectations. Moreover, knowing about the expectations from the beginning will not only help to focus the learners, it is also an important precaution to avoid disappointment or false expectations with parents and learners when it comes to grading.

Teaching and learning to the test is therefore not a negative thing, if the test assesses the communicative goals of a learning sequence and if the tasks in the test are competence oriented. However, if tests use task types that are far from being suitable to assess communicative competence, their impact on teaching and learning will be negative.

Therefore, guiding teaching and learning through goals based on criteria and descriptors that can also be used in the assessment of performances during and at the end of the learning process is considered the way forward in a culture of testing and assessment *for* learning.

References

Alderson, C. J., Clapham, C., & Wall, D. (2004). *Language Test Construction and Evaluation* (Eighth printing ed.). Cambridge: Cambridge University Press.

Allan, D. (1999). Distinctions & Dichotomies: Testing and Assessment in ELT. *FELT Newsletter, 2(1)*. (T. F. Ireland, Ed.) FELT.

BIFIE (Ed.). (2011 - 2014b). *Standardüberprüfung*. Retrieved 03 10, 2014, from https://www.bifie.at/standardueberpruefung

BIFIE (Ed.). (2011-2014a). *Kompetenzen und Modelle*. Retrieved 03 10, 2014, from https://www.bifie.at/node/49

BMBF (Ed.). (2008). *Lehrplan Lebende Fremdsprache*. Retrieved 03 20, 2014, from http://www.bmukk.gv.at/medienpool/16682/bgbl_nr_ii_210_2008.pdf

BMBF (Ed.). (2009). *Verordnung der Bundesministerin für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen StF: BGBl. II Nr. 1/2009.* Retrieved 03 10, 2014, from http://www.bmukk.gv.at/medienpool/24598/vo_bildungsstandards2013anl.pdf

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

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Studies in Langauge Testing 7 ed.). Cambridge: Cambridge University Press.

Hymes, D. H. (1972). On communicative competence. In J. Pride, & J. Holmes (Eds.), *Sociolinguistics: selected readings.* (pp. 269–293). Harmondsworth: Penguin.

Marzano, R. J. (2006). *Classroom Assessment and Grading That Work.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2010). *Formative Assessment & Standards-Based Grading.* Bloomington: Marzano Reserach Laboratory.

Mewald, C., Gassner, O., Lackenbauer, F., Brock, R., & Siller, K. (2013). *BIFIE.* Retrieved 03 10, 2014, from https://www.bifie.at/system/files/dl/TR_Speaking_130805.pdf

Richards, J. C., & Rogers, T. S. (2001). *Approaches and Methods in Language Teaching* (Vol. Second Edition). Cambridge: Cambridge University Press.

Rock, M. L., Gregg, M., Ellis, E., & Gable, R. A. (2008). REACH: A framework for differentiating classroom instruction. *Preventing School Failure, 52(2)*, pp. 31-47.

Schlichterle, B., & Weiskopf-Prantner, V. (2013). *Praxiseinblicke Englisch. Version 1.0.* (ZLS, Ed.) Retrieved 2014, from NMSvernetzung: file:///C:/Users/Claudia/Dropbox/Downloads/Praxiseinblicke%20Englisch%20final1%2013.02.2013.pdf

Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners.* New Jersey: Pearson Education.

Wiggins, G., & McTighe, J. (2005). *Understanding by Design.* Alexandria, VA: Pearson.

Wormeli, R. (2006). *Fair Isn't Always Equal. Assessing & Grading in the Differentiated Classroom.* Portland: Stenhouse Publishers.

[1] Rückwärtiges Lerndesign is the German term used for Backward Design coined by Wiggins & McTighe (2005).
[2] The term “criterion oriented” is used synonymously with “criterion-referenced” in this paper
[3] Emphasis removed by the author
[4] For a structured overview of all CEFR scales (2001) see: http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Overview_CEFRscales_EN.pdf
[5] "Practice What You Preach" is track #8 on the album “Love Songs” by Barry White
[6] The last two rows only refer to years 7 and 8 when criteria that refer to basic goals (“grundlegende Allgemeinbildung”) are defined.
[7] The last two rows only refer to years 7 and 8 when criteria that refer to basic goals (“grundlegende Allgemeinbildung”) are defined.
[8] For a thorough description of the construct of the E8 Writing Test and the Writing Assessment Scale see https://www.bifie.at/node/1497