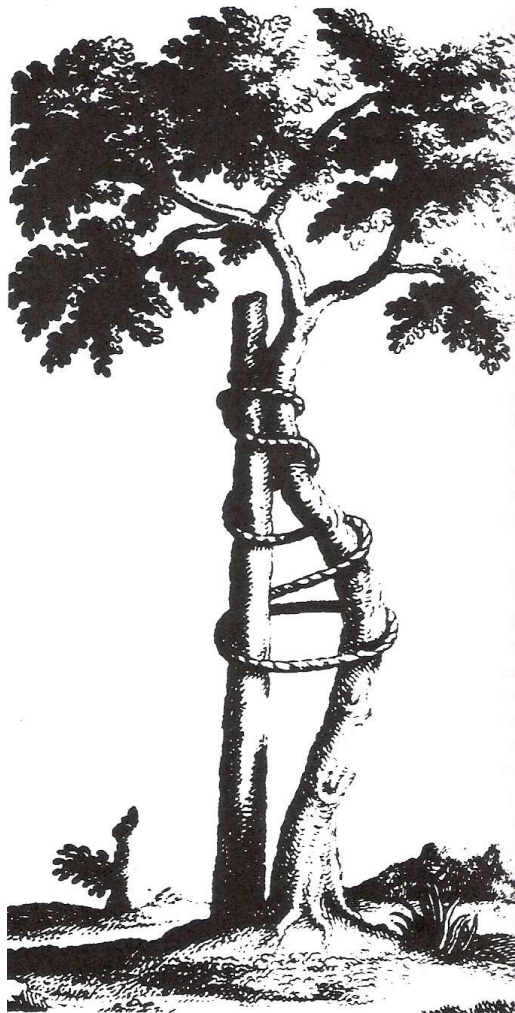


II. 1 „Mark my words.“ Bewertung in der Diskussion

Micheál Ó Dúill



N. Andry, *L'orthopédie ou l'art de prévenir et de corriger dans les enfants les difformités du corps*. 1749 (Quelle: Shohamy 2001: 56)

Wie ein Gärtner, der den Zustand seines Gartens bewerten möchte, nicht Recht hat, wenn er glaubt, dies nur anhand der reifen und Früchte tragenden Bäume bewerten zu können, obwohl er auch die noch erst reifenden Bäume berücksichtigen müsste, so muss auch der Psychologe bei der Einschätzung eines Entwicklungsstands nicht nur die reifen, sondern unbedingt auch die noch reifenden Funktionen berücksichtigen, nicht nur das aktuelle Niveau, sondern auch die Zone der nächsten Entwicklung. Wie soll das geschehen? Vygotski, *Denken und Sprechen*: 2002 (¹1934): 362

1.1 Einführung

Als ich vor kurzem Andrys Gemälde im Monograph von Shohamy entdeckte, dachte ich sofort: ‚Siehe da: Das Seil ist der Test, der Baum der arme Schüler!‘

Ich habe mich in der Zwischenzeit daran gewöhnt, dass meine Kinder solchen Vorlagen oft Anderes oder mehr entnehmen als ich. Genauso wie ich sah mein fünfzehnjähriger Sohn, ohne dass ich selbst etwas sagte, im Seil einen Test. Doch er ergänzte meine Beobachtung um zwei Kategorien: Erstens sei der Pfosten der Lehrer; zweitens aber sage das Bild aus, dass der Schüler über seinen Lehrer hinauswachse.

Meine achtzehnjährige Tochter gewann diesem Gemälde wiederum eine andere Erkenntnis ab: Der Baum sei ganz gesund gewachsen, so wie ein Baum natürlich wachse; dann sei aber ein Mensch mit einem Seil gekommen, der gemeint habe, nur ein gerader Baum sei ein richtiger Baum.

So viel zu der unterschiedlichen Apperzeption eines wissenschaftlichen Begleiters, der selbst hauptamtlich unterrichtet und prüft, und der seiner Kinder, die noch als Vollzeitschüler unterrichtet und geprüft werden.

Sowohl das Bildmotiv als auch das sprachliche Motto, die ich diesen Ausführungen als Gedankenanstöße vorangestellt habe, bedienen sich einer Metaphorik aus der Welt der Pflanzen.

So stellt der russische Psychologe Lev Vygotskij (1896-1934), dessen Einfluss in der Fremdsprachenerwerbsforschung stets zunimmt (Vgl. Johnson 2001: 182; Day 2002: 14), eine Frage, die nicht nur für das Lehren und Lernen von Fremdsprachen, sondern auch für das Testen von Fremdsprachenkenntnissen von höchster Relevanz ist: Wie soll eine Einschätzung eines Entwicklungsstandes alle relevanten Funktionen berücksichtigen? Ja, wie nur?

1.2 Definitionen

discussing testing-related issues [...] define clearly the terms used
Figueras (2005: 48)

Antragsgemäß wurde in der zweiten Phase des BLK-Modellversuchs EU-KonZert von November 2004 bis Oktober 2005 an der ‚[g]emeinsamen Standardfindung bei [der] Bewertung schriftlicher Prüfungsleistungen‘ (Staatsinstitut 2003: 13) gearbeitet.

Im deutschen Teil des mehrsprachigen Sprachtestglossars des Association of Language Testers in Europe (ALTE) steht unter ‚Bewertung‘

- a. Prozess der Zuweisung eines Punktwerts zu einer erbrachten Testleistung durch subjektive Beurteilung.
- b. Der bei der Bewertung vergebene Punktwert (ALTE members 1998: 97).

Beim Nachschlagen des entsprechenden Querverweises im englischsprachigen Teil gelangt man zu folgender Definition:

- Rating** a. The process of assigning a score to performance in a test through the exercise of judgement.
b. The scores awarded as the result of the rating process (ibid.: 159).

Im englischen Vorschlag steht also bei a. die Feststellung eines ‚exercise of judgement‘, bei der deutschen Erklärung jedoch die Beobachtung einer ‚subjektive[n] Beurteilung‘ (Hervorhebung Ó D).

Im nächsten kurz darauf in der Reihe „Studies in Language Testing“ erschienenen Band, einem einsprachigen Wörterbuch des Sprachtestens, das am Language Testing Research Centre der University of Melbourne entstanden ist, fehlt zwar der Begriff ‚rating‘, dafür wird jedoch den Möglichkeiten eines einsprachigen Glossars entsprechend um so detaillierter auf den Begriff ‚rater‘ eingegangen:

rater

Also **judge, marker, scorer**

The judge or observer who operates a **rating scale** in the **measurement** of **oral** and **written proficiency**. The **reliability** of raters depends in part on the quality of their training, the purpose of which is to ensure a high degree of comparability, both inter- and intra-rater. Since raters are human and are therefore subject to individual **biases**, close attention is paid not only to reliability, but also to analyses of rater bias.

(Davies et al. 1999: 161)

Diese Definition von Davies et al. spiegelt damit im Gegensatz zur englischsprachigen ALTE-Definition die Betonung des subjektiven Aspektes von Bewertung der deutschsprachigen ALTE-Definition wider, indem erklärt wird, dass ‚raters‘ nur ‚human‘ seien und deshalb auch zu ‚individual biases‘ neigen.

Im ALTE-Werk bezieht sich die englische Definition des Begriffes ‚bias‘ nicht direkt auf die Aufgabe eines ‚raters‘. Eine solche Verbindung könnte höchstens vom Leser als eine einzige von verschiedenen Bedeutungen in die Definition hineininterpretiert werden:

Bias A test or item can be biased if one particular section of the candidate population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured. Sources of bias may be connected with gender, age, culture, etc.

(ALTE members 1998: 136).

Dass bei dem Begriff ‚bias‘ im ALTE-Glossar eher nicht an den Davies’schen ‚rater bias‘ gedacht wird, wird jedoch deutlicher, wenn man die deutsche Erklärung und vor allem das deutsche Stichwort dazu nachschlägt:

Stichprobenverzerrung Man kann bei einem Test oder einer Aufgabe von Stichprobenverzerrung (Bias) sprechen, wenn ein bestimmter Anteil der Prüflingspopulation begünstigt oder benachteiligt ist durch bestimmte Merkmale des Tests oder der Aufgabe, die nicht Gegenstand dessen sind, was gemessen werden soll.

(ibid.: 127)

Um noch ein wenig bei dem Versuch zu verweilen, einige englischen und die deutschen Definitionen des ALTE-Werkes mit dieser einen Begriffsbestimmung des *Dictionary of language testing* zu vergleichen, sei hier abschließend darauf verwiesen, dass beim ALTE-Buch unter ‚rater‘ der Hinweis steht ‚Refer to definition for assessor‘ (ibid.: 159); der entsprechende Querverweis zu ‚rater‘ im deutschen Teil lautet dann: ‚**Einschätzung** Vergleiche Definition von Bewerter‘ (ibid.: 101). Der englische Teil bezieht sich damit auf eine Person, die eine Tätigkeit ausübt, der deutsche Teil jedoch auf die Tätigkeit selbst.

Wenn der englische ALTE-Teil nun eine Verbindung zwischen ‚rater‘ und ‚assessor‘ herstellt, dann ist es unausweichlich, den Gebrauch dieser zwei Termini genau unter die Lupe zu nehmen.

assessor Someone who assigns a score to a candidate's performance in a test, using subjective judgement to do so. Assessors are normally qualified in the relevant field, and are required to undergo a process of training and standardization. In oral testing the roles of assessor and interlocutor are sometimes distinguished. Also referred to as an examiner or rater.

Compare: interlocutor, marker
(ibid. 135)

Etwas beruhigt kann man beim entsprechenden deutschen Querverweis hierzu feststellen, dass, wenn man im englischen Teil über ‚rating‘ zum ‚assessor‘ gelangt, die Wege im Deutschen etwas überschaubarer zu sein scheinen, da Bewertung zu folgendem Begriff führt:

Bewerter Eine Person, die der Leistung eines Prüflings in einem Test einen bestimmten Punktwert zuweist, wobei eine subjektive Bewertung erforderlich ist. Bewerter sind normalerweise im entsprechenden Tätigkeitsbereich qualifiziert und müssen sich einem Trainingsprozess unterziehen sowie bestimmten Standardisierungen unterwerfen. Bei mündlichen Prüfungen haben Bewerter und Fragesteller, auch bezeichnet als Prüfer, etwas unterschiedliche Funktionen (ibid.: 97)

Man stelle fest, dass an dieser Stelle in der englischen ALTE-Version nicht mehr von ‚exercise of judgement‘ die Rede ist, sondern – durchaus im Einklang mit der deutschen Version – von ‚subjective judgement‘ gesprochen wird.

Bei Davies et al. kommt allerdings ‚assessor‘ gar nicht vor. Dafür - was die Sache für einen kurzen übersichtlichen Abriss nicht gerade erleichtert - gibt es einen Eintrag zu ‚assess‘: **assess** See **rate**. Unmittelbar danach fängt der Eintrag zum Stichwort ‚assessment‘ wie folgt an (fettgedruckte Einträge verweisen auf Querverweise):

Assessment

A term often used interchangeably with testing; but also used more broadly to encompass the gathering of language data, including test data, for the purpose of **evaluation** and making use of such instruments as **interview**, case study, questionnaire, **observation** techniques

(Davies et al. 1999: 11)

Auf eine Besprechung der hier erwähnten, nicht aber diskutierten Querverweise soll an dieser Stelle nicht deswegen verzichtet werden, weil es gilt, die Leistungen der ALTE members und der Mitarbeiter des Language Testing Research Centre der University of Melbourne zu schmälern. Ganz im Gegenteil - die hier zitierten Definitionen sind gedacht als Anregung dafür, sich damit zu beschäftigen und zu erfahren, welch schwieriges Unterfangen es ist, leitende Begriffe eines Arbeitsgegenstandes nach dem Motto ‚Define clearly the terms used‘ definieren zu wollen. Dennoch ist es für alle Bewerter wichtig, sich darüber im Klaren zu sein, dass nicht alle Begriffsbestimmungen, denen man begegnet, überall Gültigkeit haben. Dadurch ist die Beschäftigung mit unterschiedlichen Begriffsbestimmungen eine erste Voraussetzung dafür, die Beschäftigung mit der Bewertungspraxis bei den KMK-Zertifikatsprüfungen, über die hier im zweiten Teil dieser Handreichung berichtet wird, als Leser kritisch zu rezipieren.

1.3 Bewertung

Auch wenn die Beschäftigung mit Definitionen uns vor Augen geführt hat, dass im Englischen zwischen den Begriffen ‚rater‘ und ‚assessor‘ nicht immer klar unterschieden wird, so dürfte die Leitfunktion des deutschen Begriffs der zweiten Phase des KMK-Modellversuchsantrags ‚Bewertung‘ noch nachvollziehbar sein.

Nun also, bewerten: Was ist das eigentlich?

Sieht man sich die obigen Zitate noch einmal an, so fällt auf, dass in den ALTE-Zitaten zu ‚rating‘ sowohl in der deutschen als auch in der englischen Fassung wiederholt von einem ‚Punktwert‘ (‚score‘) die Rede ist.

Wie man im Anhang dieser Handreichung nachprüfen kann, ist es auch tatsächlich der Fall, dass bei dem KMK-Fremdsprachenzertifikat, das nach Bestehen der entsprechenden Prüfung vergeben wird, eine detaillierte Auflistung der erreichten Punkte pro Prüfungsteil dem Anteil der erreichbaren Punkte für diese Teile gegenübergestellt wird.

Das andere oben zitierte Sprachtestglossar (Davies et al.) spricht allerdings beim Eintrag zu ‚rater‘ nicht von einem ‚score‘, sondern von einem ‚**rating scale**‘. Dort wird man (Davies et al. 1999: 161) auf den entsprechenden Eintrag ‚**proficiency scale**‘ weiter verwiesen, bei dem es heißt:

proficiency scale *Also rating scale*

A **scale** for the description of language **proficiency** consisting of a series of constructed **levels** against which a language learner’s **performance** is judged. Like a test, a proficiency (rating) scale provides an **operational definition** of a linguistic **construct** such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or **bands** are commonly characterised in terms of what subjects can do with the language (**tasks** and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, **fluency** and **cohesion**). Proficiency scales typically consist of sub-scales for the **skills** of speaking, reading, writing and listening. [...] Scales are descriptions of groups of typically occurring behaviours; they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the test **population** and **test purpose**. **Raters** or judges are normally trained in the use of proficiency scales so as to ensure the measure’s **reliability**.

(Davies et al. 153, 154).

Liegt hier ein Widerspruch vor? Nur ein scheinbarer. Die gegenwärtige Bewertungspraxis des KMK-Fremdsprachenzertifikats verzichtet auf Noten; auf Punktzahlen verzichtet man nicht. Um zu erkennen, warum auf Noten verzichtet wird und um zu verstehen, warum trotzdem Punktzahlen angegeben werden und was diese Punktzahlen bedeuten sollen, muss man sich noch eingehender mit dem im ersten Handreichungsteil (**Aufgabenstellung**) unter dem Abschnitt **1.7 Der BLK-Modellversuch EU-KonZert** erwähnten Begriff der kriterienorientierten Bewertung beschäftigen.

Die ALTE-Mitglieder schreiben zu diesem Begriff einer kriteriumsbezogenen Orientierung:

kriteriumsorientierter Test Test, in dem die Leistung eines Prüflings im Verhältnis zu einem zuvor definierten Kriterium interpretiert wird. Es wird vor allem geprüft, ob ein bestimmtes Ziel erreicht wurde und nicht, welchen Rangplatz ein Prüfling innerhalb einer Gruppe einnimmt
(ALTE members 1998: 110).

Davies et al. halten zu diesem Thema fest:

Criterion-referenced test(ing)

Also CRT, criterion-referenced measurement, domain-referenced test

A test that 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. [...] Test **scores** report a candidate’s **ability** in relation to the crite-

tion, ie what the candidate can and cannot do, rather than comparing his/her performance with that of other candidates in the relevant **population**, such as happens in **norm-referenced tests**. **Test results** are often reported using descriptive **scales** rather than a numerical score. In contrast to norm-referenced tests the **criterion**, or **cut-score**, is set in advance. Strictly speaking, criterion-referenced tests are only concerned with whether candidates have reached a given point rather than how far above or below the criterion they may be.

(Davies et al. 1999: 38)

Sieht man sich noch einmal den Anhang dieser Handreichung an, so kann man konstatieren, dass ein KMK-Fremdsprachenzertifikat nicht nur, wie bereits erwähnt, erreichte Punktzahlen für verschiedene Prüfungsteile angibt, sondern auch die Kompetenzbereiche der Stufen aufführt, wie diese in der entsprechenden Rahmenvereinbarung beschlossen wurden. Es ist wichtig, dass ein Bewerter versteht, dass die abgedruckten Kompetenzbereiche der Rahmenvereinbarung dem ‚**criterion**‘ (Kriterium) oder ‚**cut-score**‘ (kritischer Wert) im Davies’schen Sinn entsprechen. Die Nennung der erreichten Punktzahlen für die verschiedenen Teile der Prüfung zielt hingegen auf eine Binnendifferenzierung, die aussagt, ‚how far above [...] the criterion [the candidates] may be‘. Noch einmal zur Verdeutlichung: Die Stufenbeschreibungen des Zertifikats nehmen keine solche Binnendifferenzierung vor, sondern bescheinigen lediglich das Erreichen des jeweiligen kritischen Wertes; das Vornehmen einer Binnendifferenzierung erfolgt jedoch in der Punktvergabe, ein Vornehmen allerdings, das für Davies et al. bei kriterienorientiertem Testen streng genommen nicht dazu gehört.

Um die zentrale Bedeutung dieser Art des kriterienorientierten Testens für die KMK-Zertifikatsprüfung zu verstehen, muss man dessen Behandlung im *Gemeinsamen europäischen Referenzrahmen für Sprachen (GeR)* betrachten, aus dem die KMK-Prüfung hervorgegangen ist:

Normorientierte Bewertung (oder Bezugsgruppenorientierung) bringt die Lernenden in eine Rangfolge, die Bewertung ihrer Leistungen erfolgt relativ zu den anderen Lernenden der Gruppe.

Kriteriumsorientierte Bewertung ist eine Reaktion gegen die normorientierte Bewertung. Hier wird der Lernende lediglich in Bezug auf seine Fähigkeit in dem jeweiligen Gebiet bewertet, unabhängig von der Fähigkeit der anderen Lernenden.

(Trim et al. 2001: 179)

Warum es zu dieser Art der Reaktion gekommen ist, bringt vielleicht Hughes am besten auf den Punkt:

A test which (...) is said to be norm-referenced (...) relates one candidate’s performance to that of other candidates. *We are not told directly what the student is capable of doing in the language*

(Hughes 2003: 20, Hervorhebung Ó D. Zum Unterschied zwischen norm- und kriteriumsorientierten Bewertung, vergleiche auch Brown / Hudson 2002: 2, 5).

Noch 2000 wird festgehalten, dass ‚es keine schulische Tradition des kriterienorientierten Testens und Bewertens gibt. Wir sind als Lehrkräfte an normorientiertes Arbeiten gewohnt‘ (Aigner et al. 2000: 16). Diese Handreichungsteile sind jedoch als weitere Unterstützung zur Umsetzung der kriteriumsorientierten Bewertung gedacht.

1.4 Korrekturaufgabe vollständig gelöst?

Definition der Leistung	Punkte
Die Aufgabe ist vollständig gelöst	27-30

Im Anhang des Teiles II.2 dieser Handreichung werden mit den Deskriptoren für die Bewertung schriftlicher sprachproduktiver Leistungen für die vollständige Lösung einer Aufgabe auf allen drei Stufen 27-30 Punkte vergeben. Wann verdient aber eine Korrekturleistung so viele Punkte? Dieser Abschnitt widmet sich nun einigen der Bereiche, die von der Welt der Wissenschaft bei der Erfüllung einer Korrekturaufgabe als erschwerend angesehen werden.

Ein Standardwerk der Fehleranalyse wählt als Oberbegriff den der Devianz (dieser Begriff fehlt in ALTE members, Davies et al. und dem *GeR*), um folgende Unterteilung von Fehlerarten vorzunehmen:

The clearest and most practical classification of deviance is a four-way one:

- (i) **Slips**, or alternatively lapses of the tongue or pen (...)
- (ii) **Mistakes** can only be corrected by their agent if their deviance is pointed out to him or her (...)
- (iii) **Errors** cannot be self-corrected until further relevant (to that error) input (implicit or explicit) has been provided and converted into intake by the learner.
- (iv) **Solecisms** are breaches of the rules of correctness as laid down by purists. (James 1998: 83).

Obwohl sich der Titel des James'schen Werkes sowohl auf Spracherwerb wie auch auf Sprachgebrauch bezieht (*Errors in Language Learning and Use*), dürfte sich die Verwendbarkeit der oben zitierten Kategorien, wie beim Punkt (iii) besonders ersichtlich, vor allem auf das Lehren und Lernen von Fremdsprachen beziehen, weniger aber auf ihre Verwendbarkeit beim Testen und Bewerten von Sprachkenntnissen.

Von ähnlich begrenzter Hilfe ist eine reine Feststellung zu Bewertungsarten wie die folgende: ‚Eine der zentralen Fragen ist die, ob die Bewertung (als Faktorenbündel von Bewerter/rater, Bewertungskategorien und Bewertungsmethode) eher ganzheitlich (*holistic*) oder stärker analytisch (*analytic*) durchzuführen ist‘ (Zydatiř 2002: 144). Hier bedarf es einer umfangreicheren Auseinandersetzung mit den verschiedenen Aspekten holistischer und analytischer Betrachtungsweisen. Zydatiř nimmt selbst an gleicher Stelle den Versuch einer solchen Auseinandersetzung vor, in dem er beispielsweise folgende These vertritt:

Die global-holistische Bewertung [...] vernachlässigt vor allem die Tatsache, dass eine sprachlich-kommunikative Leistung keine einheitliche Größe ist, sondern ein vielschichtiges, differenziertes Phänomen. Ein Sprecher [...] kann sich durch begriffliche Sicherheit im Wortschatz auszeichnen, aber beträchtliche (störende) Defizite [...] im Strukturegebrauch haben. Diese Aspekte sprechen für eine stärkere analytische Bewertung (*analytic marking / scoring*); d.h. spezifische sprachliche, inhaltliche und kommunikative Leistungen sollten getrennt und in abgestufter Form beurteilt werden (ebd.: 150).

Dieses Plädoyer für eine analytische Bewertung, hier bezogen auf mündliche Interaktion, kann genauso gut als Empfehlung für die Korrektur schriftlicher Leistungen verstanden werden.

Da die Begriffe ‚holistisch‘ und ‚analytisch‘ während der Workshops des Modellversuchs wiederholt intensiv diskutiert wurden, soll an dieser Stelle nun doch noch auf entsprechende Definitionen der zwei Wörterbücher der Cambridger Reihe *Studies in Language Testing* eingegangen werden. ALTE members verweisen

beim Begriff ‚holistic assessment‘ (ALTE members 1998: 164) bzw. ‚ganzheitliche Beurteilung‘ (103) auf die Definition für ‚global assessment‘ (146) bzw. ‚globale Bewertung‘:

Bewertungsmethode für Sprech- und Schreibtests. Der Bewerter vergibt eine generelle Note oder Punktzahl entsprechend seinem Gesamteindruck über die Sprachverwendung und nicht eine Anzahl unabhängiger Bewertungen von Teilaspekten des Sprachgebrauchs (104).

Ein Querverweis leitet zur Definition von ‚analytic scoring‘ (135) bzw. ‚analytische Bewertung‘:

Eine Bewertungsmethode, die bei Sprachtests verwendet werden kann, in denen es um produktive Sprachverwendung geht, beispielsweise beim Sprechen und Schreiben. Der Beurteiler nimmt eine Einschätzung mit Hilfe einer Liste spezifischer Punkte vor. So kann beispielsweise bei einem Schreibtest die analytische Skala einen Schwerpunkt setzen bei Grammatik, Wortschatz oder der Benutzung von Satzverbindungen (94).

Eine Übereinstimmung im Gebrauch der Begriffe ‚holistisch‘ und ‚analytisch‘ ist damit zwischen ALTE members und Zydatið festzustellen. Davies et al. setzen sich jedoch differenzierter mit diesen Kategorien auseinander. Zu ‚analytic scoring‘ heißt es:

A method of **subjective** scoring often used in the assessment of speaking and writing **skills**, where a separate score is awarded for each of a number of features of a **task**, as opposed to one global score (Davies et al. 1999: 7).

Anders als ALTE members und Zydatið halten Davies et al. in ihrer Definition des ‚analytic scoring‘ auch einen Kritikpunkt gegenüber dem ‚holistic scoring‘ fest:

A criticism commonly made of analytic scoring is that the focus on specified aspects of the performance may divert **raters‘** attention from its overall effect (ebd.).

‚Holistic scoring‘ wird definiert als

A type of marking procedure which is common in communicative language testing whereby **raters** judge a stretch of **discourse** (spoken or written) impressionistically according to its overall properties rather than providing separate **scores** for particular **features** of the language produced (eg **accuracy**, lexical range) (ebd. 75).

Im Gegensatz zu Zydatið und ALTE members führen Davies et al. den Begriff ‚holistic assessment‘ ein, dem sie zwei Bedeutungen zuschreiben: entweder wird der Begriff als Synonym für ‚holistic scoring‘ verwendet (ebd.); oder aber es handelt sich um einen Begriff, der beim Arbeiten mit Deskriptoren, wie dies im Modellversuch praktiziert wurde, eine Möglichkeit zu bieten scheint, einen etwaigen Konflikt zwischen rein analytischen oder holistischen Bewertungsparadigmata zu überwinden:

A general approach to assessment which involves the awarding of one or more scores to a piece of writing or an oral performance on the basis of an overall impression, as distinct from making a count of specified **features** of the discourse. The scoring procedure may involve the use of detailed descriptive statements (**band descriptors**) for each level of **performance** and for each of a number of categories or features (such as grammatical control or coherence) or simple statements such as adequate/not adequate. The term holistic assessment thus includes **analytic** as well as **holistic scoring** methods (ebd. 74).

Teil II.2 dieser Handreichung beschreibt, wie der Einsatz von Deskription zur Bewertung von schriftlichen Aufgaben in den Bereichen Mediation und Produktion erprobt wurden. Dabei wurde es im letzteren Bereich, Produktion, für angemessen

gehalten, zweiseitige Deskriptoren zu entwickeln, die eine effektive Kombination holistischer und analytischer Bewertungsvorgehensweisen gewährleisten. Im Bereich der Mediation hingegen reichen nach einhelliger Meinung einseitige Deskriptoren aus, um diese Funktion zu erfüllen. (Zu einer kombinierten holistischen und analytischen Bewertungsweise siehe auch den nächsten Abschnitt, II.1.5.)

Was zählt bei der Bewertung einer schriftlichen Schülerleistung? Wildavsky macht hierzu auf Schreibkompetenz in der eigenen Muttersprache bezogen folgende aussagekräftige Feststellung, die ohne weiteres auch für die Beurteilungskriterien von Schreibleistungen in einer Fremdsprache relevant ist: „College professors demand correct grammar from incoming students [...]. But high school teachers rank that skill last. They care how students organize their writing“ (Wildavsky 2000: 12, zitiert nach Hinkel 2002: 57). Beim hier erkannten Problem geht es darum, zu wissen, welche Erwartungshorizonte welcher Testanwender in der Bewertung in welchem Ausmaß berücksichtigt werden sollen.

Unabhängig von den Erwartungen der verschiedenen Abnehmer von Testergebnissen spielt in der Bewertung einer Korrekturleistung die Rolle der ‚rater errors‘ eine nicht unerhebliche Rolle. North führt eine Reihe solcher möglichen Korrekturefehler auf:

The classic rater errors are halo effect: transferring judgements from a global impression to categories, or between categories, central tendency: not using the top and bottom of the scale or tending to hone in on a neutral category on a questionnaire, and variation severity/leniency (North 2000: 173).

Häufiger als über Korrekturefehler spricht man in der Literatur von dem schwierigen Umgang beim Bewerten von Fehlern der Prüflinge. North selbst schreibt an anderer Stelle:

Simplistic equation of “typical errors“ with bands on a scale of proficiency on the basis of teacher instinct or committee consensus is dangerous. It distorts the measure and it has a negative wash back effect in encouraging teachers to count mistakes rather than monitor the quality of performance (ibid.: 82).

Mit ähnlichem Tenor warnt Gogolin davor, nur das zu sehen, was nicht stimmt: „Die Aufmerksamkeit von Lehrkräften bei der Prüfung von Schülerleistung richtet sich auf das Nichtvorhandene, Nichtgekonnte oder falsch Gemachte“ (Gogolin 2003: 90).

Erschwerend bei einer auf vermeintlichen Fehlern basierenden Bewertung kommt hinzu, dass trotz detaillierter Bestimmungsversuche wie der oben zitierte James'sche die Möglichkeit der Klassifizierung alles andere als eindeutig ist: „Im Rahmen von Arbeiten zur fuzzy grammar wurde deutlich, dass praktisch jede grammatische Erscheinung eine Grauzone aufweist, in der die Sprachintuitionen unklar und verschwommen werden“ (Legenhausen 2001: 48).

Überhaupt: Bewertung von Grammatik in einer Prüfung. Warum geschieht das? Und falls dies geschieht, kann man eine entsprechende Bewertung differenziert genug vornehmen? Die erste Frage beantwortet Appel mit der kritischen These „Grammatik muss gelehrt werden, weil sie geprüft wird“ (Appel 2000: 227). Eine Antwort von Quetz auf die zweite Frage könnte ihrerseits als Infragestellung der Wertstellung eines Testverfahrens, das den kommunikativen Ansatz in den Vordergrund stellen will, angesehen werden: „die generativen und kreativen Möglichkeiten grammatischer Teilsysteme können bei einer ausschließlich kommunikativen Anordnung kaum wirksam werden“ (Quetz 2003: 125).

Um nun zum oben angesprochenen Kontrast zwischen ganzheitlichen und analytischen Bewertungsverfahren zurückzukehren, so ist es keineswegs der Fall, dass holistische Ansätze uneingeschränkte allgemeine Beliebtheit genießen. Harsch findet, dass „a holistic judgement without further guidance turned out to be highly unreliable“ (Harsch: 2004: 42). Carr behauptet laut Iwashita und Grove, dass gerade bei schriftlichen Leistungen in einer Fremdsprache eine holistische Bewertung problematisch sei:

Carr (2000) [...] argues that holistic ratings are problematic in the assessment of non-native speaker writing because variations among non-native speakers' written performances are even greater than those among native speakers, due to the linguistic constraints faced by non-native speakers (Iwashita/Grove 2003: 16).

Wenn Carr von Testleistungen von Nichtmuttersprachlern spricht, wie sieht es aus mit deren Korrekturleistungen? Sind diese mit denen von Muttersprachlern zu vergleichen? Und, um eine analoge Frage zu stellen, können muttersprachliche Sprachleistungen zu einem Bewertungsmodell für die Produktion von Nichtmuttersprachlern erhoben werden?

Beim Thema Unterschiede in der Korrekturpraxis von Muttersprachlern und Nichtmuttersprachlern stellt Barron die provokative These auf, dass eine hohe Sprachkompetenz eines Fremdsprachenlerner von Muttersprachlern nicht zwangsläufig hoch geschätzt wird:

it may be that a pragmatically competent learner is viewed negatively by native speakers of the target language, as native speakers may prefer learners to act as foreigners in certain contexts and not lay claim to membership of their society (Barron 2003: 76).

Dagegen berichtet Takala über den interessanten Befund von Sundh, der in seiner Untersuchung herausfand, dass „[...] relatively higher grades were given by the Native Speakers, relatively lower grades were given by the School Teachers [...]“ (Sundh 2003: 232, zitiert nach Takala 2005).

Was die Eignung von Muttersprachenmodellen bei der Bewertung von Leistungen von Nichtmuttersprachlern betrifft, stellt North nüchtern fest:

The concept of “native speakerness“ itself as a target is flawed because it is not a tangible, homogenous concept. It gets mixed up with domain variation – people are much better at some things than at others – and with language politics, prestige varieties etc. etc. (North 2000: 55)

Bei dieser BLK-Modellversuchsphase ging es um die Bewertung von schriftlichen Prüfungsleistungen. Konkret zur Bewertung von schriftlichen Prüfungsarbeiten gibt es noch einige kritische Bemerkungen, die an dieser Stelle erwähnt werden sollen.

Hughes weist zum Beispiel auf die Gefahr hin, bei einer Bewertung von schriftlichen Leistungen der korrekten Rechtschreibung und Interpunktion zu viel Bedeutung beizumessen: „overemphasis on such mechanical features as spelling and punctuation can invalidate the scoring of written work (and so the test of writing)“ (Hughes 2003: 33).

Ähnlich äußert sich Glaboniat: „Allen Erkenntnissen und Empfehlungen der Schreibdidaktik zum Trotz bildet das Kriterium der *formalen Korrektheit* in der Praxis nach wie vor das zentrale Element in der Beurteilung von schriftlichen Texten“ (Glaboniat 1998: 22).

Alternativ zur Betonung einer formalen Korrektheit kann man nach anderen Kriterien suchen, die bewertet werden können, doch besteht Einigkeit darüber, dass dies objektiv verlaufen kann? Man nehme das Beispiel Wortschatz: „Unfortunately lexical diversity is notoriously difficult to quantify reliably“ (Malvern / Richards 2002: 87).

Doch es ist nicht nur problematisch, lexikalische Vielfalt zu beurteilen. Auch die Tatsache, dass schreiben können mehr als nur eine wie auch immer definierte und evaluierte Beherrschung von Lexis und zusätzlich noch Grammatik bedeute, müsse berücksichtigt werden: „writing ability extends beyond vocabulary and grammar to include aspects of text discourse“ (Alderson / Banerjee 2002: 95).

Die Schwierigkeiten, die bei der Bewertung einer schriftlichen Leistung bestehen, fasst Glaboniat schließlich wie folgt zusammen: „Es gilt als unbestritten, dass die Bewertung von Texten immer ein gewisses Maß an Subjektivität in sich trägt und sich Textqualitäten niemals objektiv messen und benoten lassen“ (Glaboniat 1998: 20).

Solche Beiträge sind sicherlich der Grund dafür, warum es zur Praxis eines Staatsministeriums heißen kann: „Es ist eine gute und langjährige Tradition, dass das Staatsministerium keine starren Vorschriften für [...] schriftliche Leistungserhebungen und deren Bewertung vorgibt“ (Staatsinstitut 2002: 3).

1.5 Einige Empfehlungen der Fachwelt

An dieser Stelle soll klargestellt werden, dass die wissenschaftliche Diskussion der Bewertungspraxis nicht ausschließlich aus Skepsis und Vorbehalten besteht. Bereits in den im letzten Abschnitt angeführten Zitaten wird neben einem Aufruf zur Vorsicht oft auch bereits die Tendenz einer Empfehlung erkenntlich. Im vorletzten Abschnitt dieser Einleitung sollen einige der konkreten Vorschläge einschlägiger Veröffentlichungen Erwähnung finden.

Häufig werden verschiedene Kategorien beim Prozess der Bewertung zusammen eingesetzt: „For the purpose of observing language proficiency, fluency, accuracy, and complexity were used as a preliminary, with a subsequent focus on complexity“ (Kiernan 2002: 68).

Ein solches Bestreben, durch vielseitige Betrachtung einer Leistung zu einer ausgeglichenen Bewertung zu gelangen, wird oft von dem Versuch begleitet, durch eine differenziertere Betrachtung der verschiedenen Teilaspekte Handlungsvorschläge zu erstellen. Diese Tendenz wird gerechtfertigt mit einer kritischen Distanz zur Fehlerlinguistik zugunsten einer ausführlicheren Betrachtungsweise von Leistungsmerkmalen:

Gerade im Bereich der Evaluation bedeutet die neue Fehlerlinguistik [...] eine viel zu enge Perspektive; hier ist auf die Entwicklungen des letzten Jahrzehnts und den Trend zur Interlanguage-Forschung hinzuweisen, der es nahe legt, nicht nur die fehlerhaften Stellen, sondern die Gesamtheit der lernersprachlichen Äußerungen als Beurteilungsgrundlage heranzuziehen
(Lavric 1998: 972)

Just bei einer solchen Sensibilisierung für die Bedeutung verschiedener Bewertungskategorien gelangt die kontrastive Vorgehensweise einer gekoppelten holistischen und analytischen Betrachtung zu Ehren. Es liegt nahe, dass ein holistischer Ansatz besonders geeignet ist, um die von Lavric geforderte Gesamtheit einer fremdsprachlichen Leistung zu beurteilen, während eine analytische Bewertung hingegen die Befunde dieser ersten Einschätzung ergänzt. (Vgl. Die Definition

von ‚holistic assessment‘ im vorhergehenden Abschnitt II.1.4 oben.) So verwundert es nicht, dass die Vorzüge einer kombinierten Vorgehensweise erkannt werden. Dies wird auch im folgenden Zitat aus dem Werk, welches das theoretische Gerüst der *Canadian Language Benchmarks*, erläutert:

Sometimes performance [...] is evaluated and scored holistically, based on an overall impression. In other situations detailed scoring is used, which assigns separate scores to isolated traits; such detailed scoring procedures are called analytic. *Canadian Language Benchmarks 2000* rating scales combine both types of scoring.

- First, the overall effectiveness of communication is scored, using the effectiveness criterion (holistic). [...]

- Next, quality of communication is scored, using specific criteria that focus on relevant aspects of speaking or writing performance in a given task (analytic)

(Pawlikowska-Smith 2002: 39).

Die „Canadian Language Benchmarks“ wurden als „descriptive scale of communicative proficiency in English as a Second Language“ entworfen (Pawlikowska-Smith 2000: VIII). Auch der *Gemeinsame europäische Referenzrahmen für Sprachen* sowie die KMK-Fremdsprachenzertifikatsprüfungen arbeiten mit Deskriptoren, um kommunikative Aktivitäten zu beurteilen (S. diese Handreichung, Teil 1.2; Trim et al.: 175).

Bereits 1999 machen sich Upshur und Turner ausführliche Gedanken zur Wahl geeigneter Beurteilungsskalen. Nach Auswertung der einschlägigen Fachliteratur unterscheiden sie drei verschiedene Arten von Bewertungsskalen:

‚absolute proficiency rating‘ [...] assumes [...] that there is no task effect, or that the rating scale can itself compensate for task effects [...]

‚task proficiency rating‘ [...] holds that a single rating scale will apply to all tasks of a kind (e.g., speaking tasks). Within this view, examinee abilities are estimated from ratings of their performance on a task, adjusted by the calibrated difficulty of that task.

The third notion is that of rating according to a task/scale unit. Task effects are acknowledged; it is further assumed that the qualities of discourse that reflect progressive-ability levels will differ across tasks and populations. For example, the qualities of discourse that mark competence in telemarketing will differ from those that mark competence in psychotherapy

(Upshur / Turner 1999: 89).

Warum nun an dieser Stelle so ausführlich Ergebnisse zitieren, die keinen Bezug auf den *Gemeinsamen europäischen Referenzrahmen* nehmen, geschweige denn auf die KMK-Fremdsprachenprüfung; Forschungsergebnisse, die darüber hinaus wenig Echo im darauf folgenden wissenschaftlichen Diskurs gefunden zu haben scheinen (S. jedoch Alderson / Banerjee 2002: 95)?

Der oben zitierte Aufsatz dürfte für die KMK-Fremdsprachenprüfung meiner Meinung nach aus verschiedenen Gründen interessant sein. Upshur und Turner lehnen ‚absolute proficiency rating‘ auf Grund ihrer empirischen Ergebnisse ab: „one cannot ignore systematic task effects upon performance test scores. Confidence in absolute proficiency ratings does not appear warranted“ (Upshur / Turner 1999:105). Sie betonen in ihrer Arbeit dagegen den Bezug zwischen Aufgabenstellung und Bewertung. Dabei fragen sie nach den ‚qualities of discourse‘, die bei verschiedenen Berufen gefragt sind. Ob unterschiedliche Kriterien auf der gleichen Stufe bei verschiedenen Berufen angewandt werden sollen, wurde im Modellversuch kontrovers diskutiert. Es kam hierbei jedoch zu keinem Konsens, der es er-

lauben würde, unterschiedliche Deskriptoren für verschiedene Berufe als Bewertungsgrundlage anzuwenden.

Upshur und Turner dokumentieren ferner die zunehmende Bedeutung der Kalibrierung von Aufgaben in der gegenwärtigen Sprachtestpraxis. Der Begriff der Kalibrierung wird in die einschlägigen Glossare beziehungsweise Wörterbücher bereits Ende der 1990er aufgenommen:

Eichung Der Prozeß der Skalenbestimmung eines Tests

(ALTE Members 1998: 100)

calibration The process of determining the scale of a test or tests (ebd. 137)

calibration

The calibration of a test involves the determination of the value of **test items** against a particular measurement scale, in other words it reflects **item difficulty**.

(Davies et al. 1999: 18)

Dem Vorhandensein dieser Definitionen in zwei Bänden der international einflussreichen Reihe *Studies in Language Testing* zum Trotz scheint es nicht, als ob der Wichtigkeit der Kalibrierung von Items bis heute unbedingt viel Aufmerksamkeit geschenkt wurde: Gehe ich zumindest von einem Griff in meinen Handapparat von kürzlich angeschaffter Fachliteratur aus und kontrolliere ich bei denjenigen neueren Veröffentlichungen, bei denen ein Sachwortverzeichnis vorliegt, so stelle ich fest, dass der Begriff ‚Kalibrierung‘ in einer Reihe von Veröffentlichungen im Stichwortverzeichnis gar nicht vorkommt (Alderson 2005, 284 Seiten; Ellis / Barkhuizen 2005, 404 Seiten; Hinkel 2005, 1144 Seiten; Weir 2005, 301 Seiten; Zydatiß 2005, 402 Seiten).

Die erste Phase des Modellversuchs beschäftigte sich mit der gemeinsamen Zuordnung von Aufgabentypen zu den drei Stufen der KMK-Fremdsprachenprüfungen. Upshur und Turner lehnen in ihrer Arbeit die Anwendung einer einzigen Skala zur Bewertung unterschiedlicher Aufgaben ab:

[T]he differences between the salient qualities of discourse that emerged in the construction of the scales [...] appear to be related to the predictability of speaker intentions in the tasks (Upshur / Turner 1999: 105).

Sie plädieren stattdessen dafür, dass ‚rating scales‘ ‚not just population-specific‘, sondern auch ‚task-specific‘ sein sollten (ebd.). Sie sprechen sich in anderen Worten dafür aus, dass Skalen nicht nur pro Berufsgruppe, sondern auch pro individuelle Aufgabe erstellt werden. Obwohl sie also zu Recht auf den Einfluss der Einzelaufgabe auf die zu erzielende Leistung und auf die Bedeutung der Kalibrierung von Aufgaben hinweisen, werden ihre Vorschläge zu einer solch umfangreichen Erstellung von Skalen wohl kaum Anhänger finden können. Zudem gibt es etwa zu bedenken, dass sie als Ergebnis ihrer empirischen Untersuchung verschiedener mündlicher Testaufgaben die Tonbandaufnahme eines selbst entworfenen Briefes durch den Prüfling als vorzuziehende Aufgabe einer mündlichen Prüfung empfehlen.

Pawlikowska-Smith, aber nicht nur sie, spricht von Effektivität. In der Tat soll an dieser Stelle betont werden, dass Fachbeiträge wiederholt auf die Wirkung eines Kommunikationsbeitrages hinweisen. Bei Legenhausen heißt das dann beispielsweise: „Die Fehlerkorrektur in schriftlichen Arbeiten erweist sich in all den Fällen als problematisch, in denen die Mitteilungsabsicht des Schülers nicht eindeutig rekonstruiert werden kann“ (Legenhausen 2001: 48).

Buck seinerseits nimmt eindeutig Stellung: „testers should be less concerned with how much a person knows about the language, and more about whether they can use it to communicate effectively“ (Buck 2001: 83).

Geht es um eine treffende Einschätzung der Effektivität der Leistung eines Prüflings, darf aber vor allem nicht vergessen werden, wie sich Muttersprachler in der entsprechenden Situationen tatsächlich verhalten. In anderen Worten „sollten der Stellenwert und die Gewichtung von sprachlicher Korrektheit auch in Prüfungen nicht höher angesetzt werden als in realen Situationen und in einem ausgegogenen Verhältnis zu anderen Kriterien stehen“ (Glaboniat 1998: 22-23).

Ein solches Betrachten tatsächlichen Gebrauchs spannt notwendigerweise den Bogen zur erwarteten situativen Einbindung der Leistung des Prüflings, denn sobald eine solche Herausforderung an den Prüfling gestellt wird, muss es selbstverständlich sein, dass nicht nur bei der Aufgabenerstellung, sondern auch bei der Bewertung der Prüfungsleistung ein realitätsnaher, fachrelevanter Erwartungshorizont ausschlaggebend ist:

Just as we analyse the target language use situation in order to develop the test content and methods, we should exploit that source when we develop the assessment criteria. This might help us to avoid expecting a perfection of the test taker that is not manifested in authentic performances in the target use situation (Alderson / Banerjee 2001: 224).

Mit dieser Übersicht einiger einschlägigen Veröffentlichungen zur Bewertung sprachlicher Leistungen habe ich den Versuch unternommen, die Ergebnisse der zweiten Phase des Modellversuchs, die hier im Teil II.2 dokumentiert werden, in den weiteren Zusammenhang der gegenwärtigen Entwicklungen im Bereich des Sprachtestens zu stellen.

1.6 Envoi

Quis custodiet ipsos custodes
Juvenal, *Satura VI*, 347-348

Jeden Sommer nach Abschluss der Unterrichts-, Prüfungs- und Korrekturzeit packe ich mit meiner Familie unsere sieben Sachen zusammen und wir verschwinden in den Urlaub. Wir erobern die Buchhandlungen am Urlaubsort und lesen bis tief in die Nacht vorm offenen Kaminfeuer all das, was der Buchhandel uns geschenkt hat. Alle Tätigkeiten des sonstigen Alltags werden ausgeblendet, keine einzige köstliche Urlaubssekunde wird der Problematik der gerechten Bewertung gewidmet.

Keine einzige?

Als ich in diesem Sommer wie versessen in *Digital Fortress* von Dan Brown las – ich hatte es schon fast fertig gelesen, alle anderen in der Familie schliefen schon, der Nebel drang bereits unter die Haustür hindurch – da stieß ich auf dieses Zitat von Juvenal:

[...] “Quis custodiet ipsos custodes.” It translates roughly to –
‘Who will guard the guards!’ Susan interrupted.
(Brown 1998: 465).

„Wer wird die Wächter bewachen?“ Plötzlich war ich wieder voll im BLK-Modellversuch EU-KonZert. (Doch nicht nur mich treibt dieses Juvenalzitat um: Seit ich den ersten Entwurf dieses Handreichungsteiles verfasste, stellte Charles Alderson auf das Sprachtesten bezogen die gleiche Frage [Alderson 2006: 59]).

Die Wächterfunktion:

“High-stakes tests [...] The primary use of such tests is “to ration future opportunity, as the basis of determining admission to the next layer of education or to employment opportunities“ (Chapman / Snyder 2000: 458)
(Andrews 2004: 37)

Man könnte an dieser Stelle die Meinung vertreten, dass es sich bei der KMK-Fremdsprachenzertifikatsprüfung um keinen ‚high-stakes test‘ handelt, da bei einer freiwilligen Prüfung nicht viel auf dem Spiel steht. Dem entgegen halten müsste man als Vertreter dieser Prüfung die im Handreichungsteil I. 1 zitierte Feststellung der Forscher des Instituts der deutsche Wirtschaft Köln, dass „[b]ei der Einstellung von Mitarbeitern [...] die Mehrzahl [...] der Unternehmen auf Sprachzertifikate“ achte (Schöpfer-Grabe / Weiß 1998: 155).

Dennoch ist es so, dass der BLK-Modellversuch EU-KonZert darauf aufbaut, ein Evaluationskonzept zur Sicherung der Vergleichbarkeit der Standards der KMK-Fremdsprachenzertifikatsprüfungen zu entwickeln. Dies erfolgt, indem länderübergreifend das im Zertifikatssystem tätige Personal unterstützt wird (vgl. Staatsinstitut 2003: 8). Eines der Kernanliegen des Modellversuchs ist es, dass das während des Projektes aufgegriffene Expertenwissen nachhaltig verfügbar gemacht wird (ibid., 2).

Aus diesen Gründen kann man zuversichtlich behaupten, dass auch bei der Gestaltung und Bewertung der entsprechenden Prüfungen die an den KMK-Fremdsprachenzertifikatsprüfungen Beteiligten ihre Wächterfunktion erfüllen, ohne dass die Wächter eines Wächters bedürften.

Gerade deswegen sei es erlaubt, an dieser Stelle abschließend darauf hinzuweisen, dass die KMK-Prüfung keine Angelegenheit ist, die im gesellschaftsfreien Raum stattfindet. Fulcher drückt das so aus: „the ultimate challenge is how we utilise the range of tools at our disposal to address questions that are essentially human and social in essence“ (Fulcher 2003: xii).

Man kann es auch anders ausdrücken, in einer Weise, die dem den BLK-Modellversuch EU-KonZert kennzeichnenden Professionalismus und Engagement, nicht abspricht, die aber daran erinnert, dass das Prüfen immer in einem größeren Kontext gesehen und nun doch selbst bewacht werden muss: „There is that in learning and thinking, in the teaching-learning relationship, in the process of education and being educated, which is immeasurable“ (Carini 2001: 174).