III.1 "Now you’re talking!" Überlegungen zu mündlichen Sprachprüfungen

Micheál Ó Dúill

1.1 "Oh, now I come? Yes, …"

Fangen wir mit Ausschnitten von drei Transkripten aufgezeichneter mündlicher Prüfungen an. Auf welcher Stufe fanden diese Prüfungen wohl statt? Und wie sind die jeweiligen Leistungen der Prüflinge zu bewerten? (Erste Rohtranskripte dieser Aufnahmen wurden von Maike Engelhardt erstellt.)

Beispiel 1

A: Good afternoon.
My name is [A.A.].
I’m, you’re on the hotline from Western Tech.
What can I do for you?

B.: Oh, hello, my name is [B.].
Eh, about one month ago I bought a CD Rom eh at West-, Western Tech in the internet.
And my reason for this phone call is I have em a lot of problems with this CD Rom.

A.: Uhm.

B.: Can you help me here?

A: Um. Yes, I can help you.
Em, have you got a customer number from our firm?

B.: Yes, my customer number is 78448

A.: Oh. Okay.
Eh, can you describe your problem?

B.: Em, yes.
Eh, I put the CD Rom X 4 hundred 7 like the, the
[iks]
[EXHALES LOUDLY]
directory in my manual in the computer,
and tried to INstall the program on the disk which was in the car-, cartoon
eh where the CD Rom was
and then cames the arrow, error
[THREE-SECOND PAUSE]
CD Rom not found.
I don’t know what I shall do?

A.: Um. Em what for a ver-, version did you have about the installation program.

B.: Eh, I have the program disk utilities 4.1, second version

A.: Ah yes, eh with this program and version eh we has many problems.
But it is easy to repair this.
Em, have you got a internet connection?

B.: Yes, I have a internet connection

A.: Oh, good,
em,
you go in the internet on the site tribble w dot western tech dot com.
And you go on the link to downloads.
You download the file bug fix d u 4 point 1
and after these you have downloaded the file you installed these program
After the installation you em start the computer at new and then em
the problem don’t, don’t were on your computer.

B.: Ok, I’ll try to download this program and inSTALL it and then we will see.

A.: Ok

B: Eh, thank you for your help, and bye-bye

A.: Thank you.
A very nice day.

Beispiel 2

C: Hello.
Em.
My name is [C].
Eh, I’m a trainee,
Em
[3 SECOND PAUSE; LOOKS AT NOTES]
I’m a trainee em [1] from the United States and eh I’m here to help you.
Em. [1]
Or can I help you
[DROPPING INTONATION]

D: Yes, of course.
Hello, my name’s [D.D.]
I’m a council clerk from this administrative community.
I’m responsible for the move and I’m VERY glad of your help.

C: Em, well, em, can you say me where the new department are?

D: Em, the new depar-, eh, em,
the new department is em,
the new financial administration

C: Uhm

D. is now in the Hauptstraße at the first floor.
Em, have you ever been in that new building?

C: Yes, eh, I know.
I was there eh in the past
Em.
At the last week I was eh in the building control office.
Eh, it’s the second, second floor. Ém
To, to, to, to, eh, to translate a fax, facsimilimie message or so
[FOUR-SECOND PAUSE
D. LOOKS AT NOTES]

D: Can you bring some files to the financial adminstration for me please
and can you order there in the shelf, please?
[TWO-SECOND PAUSE]

C: Eh, yes, I can eh take this files em to the financial administration.
Em, is there, is it em next to the mail room eh
or opposite em to the personal department

D: You must go upstairs to the first floor,
then turn left and go till the end of the floor.
There is the financial administration on your right hand side
and next to mailroom.

C: Yes, thank you.
Eh, and em can I enter the data into the computer or take copies for you em?
And eh the computer, at the computer, is there an internet access?

D: The new computer has an internet access, yes.

C: Very well

D: With a printer, a mouse

C: Oh

D: And a keyboard.
Do you come to the meeting at eh three o’clock in the AFTERnoon, too?

C: Sorry, em, I can’t eh come to the meeting em before half past three.
Em I have a me, I have em a meeting em with the mayor em
[THREE-SECOND PAUSE]
Em so em
Now em I will take this file and em bring to, bring to the financial administration.

D: Thank you, very kind of you.
Em, thank you for your help, and we see us.
C: No thanking, it’s my job.
D: OK
Bye
C: Bye
[C. AND D. SHAKE HANDS]

Beispiel 3

E: Well, I think eh Europe PasSAGE,
F: Umhm
E: eh maybe we eh have a little focus on Europe
So we have a lot of eh Europe Asian food, Europe mediterrean food,
but eh I think eh, my mind is coming now, eh
Polynesia,
never, never eh heard of eh Polynesian restaurants
so I don’t know what kind of food they have, so
F: But you know it doesn’t belong to Europe?
E: Yeah. So Europe. Exactly. So that I,
then I thought em,
we can do a European Polynesian eh, eh, eh restaurant.
F: Wow
E: S-, something new. Never, never works
G: I mean downtown Hamburg here
F: OK
G: there’s a lot of business business
F: you can have it in the traders’ X
F: as well. Polynesian, Polynesian and European food
E: There’s Polynesian? Then let me take: New,
E: New Zealand? Oder Madagaskar
F: [EXHALES AIR]
E: there is the movie, the movie from the Walt Disney,
Madagaskar, with the yeah
F: Umhm.
E: Yeah
F: Ok
E: Yeah. The animation, the animation movie. Europe.
F: But I think it would be a good idea to
to have the focus on the European, yes,
G: just European?
I mean it’s a lot a lot of business in town, right, in the Europa Passage,
so a lot of offices and everything, so, em,
maybe we should do something which is kind of fast
and kind of eh good, fresh and everything
so whatever that the people come there for for lunch
then we really specialize on, on the lunch em time.
Like the Vapiano thing, which is already sold in pasta,
maybe we can do it with something else.
some, maybe Polonesisches, maybe one never know.
What-, whatever.
But still that it’s fast and [WHISTLES] goes out and;
I guess it’s a very good place to, to open a restaurant.
E: well
G: To make money
E: The buffet, the buffet style is eh, eh, eh, eh fast
But eh it’s, it’s you know like a, like a modern kind of CANteen.
Go in, go out.

43 But not the canteen you know from eh from eh
44 from a business like Mercedes Benz or whatever.
45 Something more in a rest'ran'.
46 A mixture, mixture, mixture.
47 What do you think about that?
48 G: Yea. Or an American thing, like at the Hooters thing, you know?
49 E: and then you can
F: yes!
50 E: then you can use all the European XXX
[F. & G. LAUGH]
51 G: I mean sex sells, still, you know. So, em
52 F: It does
53 G: So we just
F: OK
54 G: girls with small pants on and you know
55 E: And rollerskates?
56 F: And rollerskates
G: Yeah, yeah
57 G: so there you go
F: rollerskates
58 G: Yes, and that's fast as well. 'Cos they can
59 F: OK.

Welche Stufen?

Und wie viele Punkte pro Prüfling?

Mehr hierzu in Abschnitt 1. 6 unten.

1.2 Einige weitere Definitionen

Wir werden diesen dritten Teil der Handreichung, der sich mündlichen Prüfungen widmet, nicht schon wieder mit Definitionen beginnen? Wurde doch etwa bei Teil I. 1.6 oben der Begriff des „Testkonstrukts" ausführlich erläutert. Aber nicht nur im Teil II. 1.2 („Definitionen"), sondern auch im anschließenden Teil II. 1.3 („Bewertung") detailliert auf Begriffsbestimmungen eingegangen, die für die Bewertung von schriftlichen Prüfungsleistungen bedeutend sind.

Will man sich mit der These auseinandersetzen, mündliche Sprachproduktion sei grundsätzlich zu unterscheiden vom schriftlichen Sprachgebrauch, ob bei Muttersprachlern oder bei Lernenden, dann kann man wohl nicht umhin, sich mit den Beschreibungsversuchen zu beschäftigen, die versuchen, diese Andersartigkeit mündlicher Äußerungen zu erkennen und festzuhalten, denn diese Andersartigkeit, soweit es sie tatsächlich geben sollte, müsste von größter Relevanz bei der Entwicklung eines Testkonstrukts sein, der als Grundlage für das Prüfen mündlicher Fertigkeiten dienen soll.

Das englische Wort ‚proficiency' kam in einer Reihe von Zitaten in Teilen I. 1 und II. 1 hier vor, ohne dass jedoch eine Begriffsbestimmung vorgenommen worden sei.

Was bedeutet nun eigentlich dieses Wort?

Im *Multilingual glossary of language testing terms* heißt es im englischsprachigen Glossarteil:

> **Proficiency** Knowledge of a language and degree of skill in using it
> (ALTE members 1998: 158).

Im deutschsprachigen Teil hingegen etwas ausführlicher:

> **Beherrschung** Kenntnis einer Sprache und Fertigkeit bei ihrer Anwendung sowohl in schriftlicher als auch mündlicher Form
> (ebd.: 96)

Wie wir bei Teil II. 1.2 oben gesehen haben, gibt das ein Jahr später in der gleichen Reihe „Studies in Language Testing“ der Cambridge University Press erschienene Werk *Dictionary of language testing* gern etwas ausführlichere Definitionen an:

> **Proficiency**
>
> There are three main uses of the term **proficiency**.
>
> 1. A general type of **knowledge** of or **competence** in the use of a language, regardless of how, where or under what conditions it has been acquired;
> 2. **Ability** to do something specific in the language, for example proficiency in English to study in higher education in the UK, proficiency to work as a foreign language teacher of a particular language in the United States, proficiency in Japanese to act as a tour guide in Australia.
> 3. **Performance** as measured by a particular testing procedure.
>
> (Davies et al.: 1999: 153)

Spricht ALTE members 1998 von „Knowledge“ bzw. „Kenntnis“, hilft uns die dritte Definition von Davies und Kolleginnen und Kollegen noch weiter, indem sie von „Performance“ („Performanz“, ALTE members 1998: 117) spricht als das, was in Tests gemessen wird. Wir haben bereits in den Teilen I. 1 und II. 1 gesehen, wie der Begriff „Performanz“ mit der Verwendung von Deskriptoren und Stufenbeschreibungen zur Feststellung des Grades der Sprachbeherrschung wiederholt in Verbindung gebracht werden, beispielsweise bei Brindley 1998, Davies et al. 1999, North 2000 und Morrow 2004. North ist bekanntlich derjenige, dessen Forschungen für das empirische Gewicht des von ihm mitverfassten *Gemeinsamen europäischen Referenzrahmens für Sprachen* (Trim et al. 2001) von zentraler Bedeutung sind. In seinem einschlägigen Werk zur Entwicklung dieses Referenzrahmens bestätigt nun auch er einen Trend im Gebrauch des Begriffs „proficiency“, der wie bei der soeben zitierten dritten Definition von Davies et al. testtheoretisch auf eine Subsumierung der Begriffe „competence“ und „ability“ hindeutet:

> The term proficiency is preferred to competence in the context of the current study because there is such confusion over whether or not the concept of ability can be included in competence
> (North 2000: 43).

An dieser Stelle sei darauf hingewiesen, dass der von ALTE members vorgeschlagene Begriff „Beherrschung“ nicht von allen als Übersetzung des englischen Wortes „proficiency“ bevorzugt wird. So spricht etwa Zydatiß (2002: 89) von der „Sprachfähigkeit in einer Fremdsprache […] (=*proficiency*) […] das Kompetenzniveau eines Fremd- (oder Zweit-)sprachenlerners zu einem bestimmten Zeitpunkt einer Sprachbiographie“.

Da wir uns in diesem Handreichungsteil mit dem mündlichen Sprachgebrauch beschäftigen, wollen wir uns nun, nachdem wir gerade den in den ersten beiden Handreichungsteilen erwähnten Begriff des „proficiency“ wieder aufgegriffen und genauer erläutert haben, dem verwandten Begriff des „oral proficiency“ widmen.

ALTE führt den Begriff „oral proficiency“ auf und definiert diesen so: „Competence in speaking a language“ (ALTE members 1998: 155). Als deutsches Äquivalent wird der Terminus „Sprechfertigkeit“ empfohlen und definiert als „Kompetenz beim Sprechen einer Sprache“ (ebd.: 125). Davies et al. 1999 führen einen Begriff „oral proficiency“ in ihrem Wörterbuch nicht auf.

Bei „oral proficiency“ geht es also bei ALTE nicht mehr, wie oben zitiert bei der Definition von „proficiency“ um „knowledge of a language and degree of skill in using it“, sondern wieder um „competence“. Vielleicht sollten wir uns näher ansehen, wie der Begriff „competence“ in diesem mehrsprachigen Werk sonst verwendet wird.

Dazu vorab einige kurze Bemerkungen zur prägenden Präsenz unterschiedlicher Sprachen im gegenwärtigen Wissenschaftsdiskurs.

Bereits die angegebene Autorenschaft dieses für alle, die sich in den zehn darin verwendeten Sprachen mit dem Sprachtesten beschäftigen, einmaligen Werkes macht deutlich, dass es ein Produkt mehrerer Hände und Tastaturen ist. Weiteres halten die zwei Seiten „Acknowledgements“ (ebd.: IX, X) fest, die etwas Aufschluss darüber enthalten, ob man bei den verschiedenen Sprachteilen von „Äquivalenten“ oder „Übersetzungen“ reden sollte. Eine eindeutige Feststellung erlauben diese „Acknowledgements“ jedoch leider nicht, obwohl dies so wichtig wäre, um den Einfluss dieses Werkes bei der Prägung des entsprechenden transnationalen testtheoretischen Sprachgebrauches zu überprüfen. So werden beim Eintrag zum deutschsprachigen Teil eine Vertreterin des deutschen Goethe-Instituts und ein Vertreter des deutschen Bundesinstituts für Berufsbildung als Übersetzer aufgeführt (ebd.: IX); beim entsprechenden Eintrag zum englischsprachigen Teil wird jedoch niemand als Übersetzer genannt.

Bei der trotz der Hervorhebung des Begriffs „proficiency“ etwas gespenstisch stets zurückkehrenden Bezeichnung „competence“ heißt es bei ALTE members im entsprechenden Haupteintrag in folgender Ausführlichkeit (1998: 138):

> The knowledge or ability to do something. Used in linguistics to refer to an underlying ability, as contrasted with performance, which is the manifestation of competence as language in use. This distinction originates in the work of Chomsky.

Hierzu zur Ergänzung die deutschsprachige Fassung, die den Begriff interessanterweise leicht anders definiert (ebd.: 108):

> Wissen oder Fähigkeit, um etwas zu tun. In der Linguistik bezeichnet Kompetenz eine zugrunde liegende Fähigkeit im Gegensatz zur beobachtbaren Performanz, die sich in der Sprachverwendung als Ausdruck der Kompetenz niederschlägt. Diese Unterscheidung stammt aus den Arbeiten von Chomsky.

Davies et al. bieten keine Definition von ‚competence’ als Oberbegriff an, sondern verweisen auf spezifische Kompetenzen:

> **Competence**
>
> *See* **linguistic competence, discourse competence, pragmatic competence, strategic competence**
> (Davies et al. 1999: 27)

Dieses Hervorheben von einzelnen Kompetenzen erinnert uns an die in Abschnitt II.1.5 oben zitierte Bemerkung von Upshur und Turner, dass „the qualities of discourse that mark competence in telemarketing will differ from those that mark competence in psychotherapy“ (Upshur / Turner 1999: 89). Diese Feststellung zu Diskurskompetenzen wurde in einem Aufsatz zu „speaking ability“ gemacht. Um die-

sen Abschnitt abzuschließen, möchte ich nun in einiger Ausführlichkeit eine etwas ältere Abhandlung zu „oral proficiency“ zitieren, die bis heute nichts von ihrer Relevanz eingebüßt hat. Diese Relevanz ist um so wichtiger, weil sich die folgende Bestimmung des Begriffs „oral proficiency“ in ähnlichem Maß auf alle die in den drei folgenden Abschnitten zu besprechenden Aspekte mündlicher Prüfungen, die Aufgabenstellung, Prüfungsdurchführung sowie Bewertung bezieht:

> The preconditions for a display of oral proficiency […] include […]:
>
> 1. Face-to-face interaction […]
> 2. Decision-making opportunities. […] [C]hoice of *when* to speak, for *how long* and *about what*.
> 3. Goal relatedness. […] [L]inguistic (lexical, syntactic, prosodic, etc.) tools used in the interaction are not the only yardstick for the evaluation of quality
>
> (van Lier 1989: 493-494)

1.3 Aufgabenstellung

Im Folgenden werde ich mich auf einige Veröffentlichungen beziehen, die sich speziell mündlichen Prüfungen widmen.

Im soeben zitierten Aufsatz von van Lier betont der Autor die Bedeutung von Interaktion, Wahlmöglichkeiten im Gespräch sowie Zielorientierung als Charakteristika mündlicher Prüfungen. Im weiteren Verlauf seines Artikels entwickelt van Lier seine Thesen von den Eigenschaften mündlicher Kommunikation. Wie diese in mündlichen Prüfungen wie dem so genannten „oral proficiency interview“ getestet werden, steht für ihn eindeutig die Fähigkeit, Konversationen zu führen (und nicht etwa Vorlesungen zu halten) im Vordergrund und deswegen ergänzt er seine Thesen zu „oral proficiency“ mit folgenden Merkmalen der Konversation: „face-to-face interaction, unplannedness (locally assembled), unpredictability of sequence and outcome, potentially equal distribution of rights and duties in talk“ (ebd.: 495).

Erkennt man in diesem zweiten Zitat von van Lier wieder die Hervorhebung der Bedeutung echter Kommunikation (Interaktion und nicht etwa das Reagieren auf von Tonträgern vorgespielte Aufnahmen), so wird hier ebenfalls betont, dass eine Interaktion so gestaltet werden muss, dass deren Verlauf nicht von nur einem der Teilnehmer (etwa einem Prüfer oder einem selbstbewussten Prüfling) dominiert wird, sondern dass beide bzw. alle Beteiligten Gelegenheit haben sollten, die Entwicklung einer Unterhaltung entscheidend mitzuprägen. Allerdings scheint ein potentieller Widerspruch zu bestehen zwischen dem Wunsch nach „goal relatedness“ und „unpredictability of sequence and outcome“. Welche Aufgabenstellung könnte denn beides gewährleisten?

Ein Ausweg aus diesem scheinbaren Dilemma könnte van Liers Betonen der Bedeutung des Kontexts beinhalten, die, soweit bei der Aufgabenstellung berücksichtigt, wohl eine Möglichkeit einer gewissen Beurteilung der Realisierung des Ziels des so genannten „goal-relatedness“ bieten könnte:

> Every contextual manifestation of speaking ability requires, in addition to speaking ability, context-specific skills and experience. Indeed, the fact that speaking ability does not exist in isolation from contexts means that every test of speaking is a performance test
> (ebd.: 500).

Hier denkt man an die im letzten Abschnitt zitierte Gleichstellung von “proficiency” und “performance” (Davies et al. 1999: 153) sowie an die im Abschnitt I. 1.4 oben

zitierte Feststellung, dass „every attempt should be made to allow background knowledge to facilitate performance rather than its absence to inhibit" (Alderson 2000: 121). Eine sorgfältige Überlegung des gewünschten Kontexts bei der Aufgabenstellung wird in anderen Worten dazu beitragen, dass die Prüflingsperformanz adäquat geprüft wird; darüber hinaus gewährleistet die Akzeptanz des Hintergrundwissens des Prüflings die Testvalidität, das heißt, dass das, was geprüft werden soll, tatsächlich geprüft wird. Zur Erinnerung:

> Ein Test oder ein Beurteilungsverfahren kann in dem Maß ‚valide' genannt werden, in dem man nachweisen kann, dass das tatsächlich gemessene Konstrukt auch das ist, das in dem betreffenden Kontext gemessen werden soll
> (Trim et al. 2001: 172; Vergleiche Abschnitt I. 1.6 oben).

In wie weit spiegeln sich die Analysen van Liers von vor gut anderthalb Jahrzehnten mit den Beobachtungen zeitgenössischer Forscher wider?

Hardy und Moore fassen die Ergebnisse einer Untersuchung aus dem Jahr 2001 von Robinson wie folgt zusammen:

> A complex task (with unfamiliar content and greater processing demands) led to greater speaker-listener interaction by inducing comprehension checks on the part of the hearer and greater lexical variety and a lower number of words on the part of the speaker
> (Hardy / Moore 2004: 343).

An diesem Zitat erkennt man, wie sich die Wissenschaft seit van Lier bemüht, zu untersuchen, welche Art von Aufgabenstellungen für eine reiche mündliche Interaktion förderlich ist. Nach Robinson (2001) besteht beispielsweise ein Zusammenhang zwischen der Komplexität der Aufgabenstellung und der lexikalischen Vielfalt der Teilnehmer: Es wird weniger gesprochen, dafür benutzt man einen größeren Wortschatz, um die Komplexität der zu äußernden Inhalte auszudrücken.

Den bisher erwähnten Forschungsergebnissen und Empfehlungen etwa von van Lier, Alderson und Robinson gemäß scheint die Annahme berechtigt zu sein, dass eine komplexe mündliche Aufgabenstellung, die den Kontext des zu prüfenden Konstrukts berücksichtigt, die sei, die zu empfehlen ist. Doch wie steht es mit der leidigen Frage der Stellung der so genannten „languages for specific purposes" (Vergleiche hierzu abschnitt I. 1.4 oben)? Lumley und O'Sullivan berufen sich auf Davies, der schreibt, dass „sampling for a proficiency test should not be restricted to the work domain" (Davies 2001: 138, zitiert nach Lumley / O'Sullivan 2005: 417); diese Wissenschaftler zeigen sich überzeugt von der Davies'schen These, dass „test tasks relevant to a broader social domain" Teil jeder mündlichen Prüfung zu sein haben (Lumley / O'Sullivan ebd.).

Neben den Empfehlungen, mündliche Interaktion kontext- aber nicht ausschließlich berufsbezogen sowie durch komplexe Aufgabenstellungen zu prüfen, gibt es auch Stellungnahmen zur geeigneten Anzahl der zu stellenden Aufgaben. Lee bemerkt hierzu unter Berücksichtigung der durchführungs- und bewertungstechnisch relevanten Frage der Anzahl der Bewerter (zur Rolle der „raters", siehe oben II. 1. 2 Definitionen und vor allem II. 1.3 Bewertung):

> Univariate analyses have shown that, to maximize score reliability for speaking, it would be more efficient to increase the number of tasks than the number of ratings per speech sample. […] While tasks are on average comparable in difficulty, they are not uniformly difficult for all examinees. […] However, beyond five or six tasks, there would be clearly a diminishing return in increasing the number of tasks
> (Lee 2006: 162).

Wichtig an dieser Feststellung aus der Sichtweite der anzustrebenden Testfairness ist die Erkenntnis, dass die Tatsache, dass nicht jeder Prüfling jede Aufgabe ähnlich gut bewältigt, eine empirisch objektiv festzustellende Tatsache darstellt, die eher durch eine anzustrebende Pluralität der Aufgabenstellungsformen zu berücksichtigen ist denn durch exzessiv strenge Bewertung zahlenmäßig begrenzter Aufgabenformen. Die im vorangehenden Abschnitt wieder zitierten, vor allem aber im Abschnitt II. 1. 5 ausführlich besprochenen Abhandlung von Upshur und Turner beschäftigt sich eingehend mit dem Umgang mit den so genannten „test effects“ in verschiedenen Bewertungsvorgehensweisen.

Wenn in diesem Abschnitt zunächst die wegweisenden Beobachtungen van Liers betont wurden und dann über die Beleuchtung einzelner Aspekte mündlicher Prüfungen durch spätere Wissenschaftler berichtet wurde, erfolgt dies nicht, um den Eindruck zu erwecken, dass sich seit van Lier keiner an eine größere Darstellung der Charakteristika mündlicher Prüfungen gewagt hätte. Es ist in der Tat so, dass zum Beispiel Fulcher sehr ausführliche Ausführungen zu diesem Thema vorgelegt hat. Wegen ihrer Qualität, Relevanz und Detailliertheit verdienen es diese, hier auch ausführlich zitiert zu werden. So entwirft Fulcher einen ganzseitigen Rahmen für das Testkonstrukt einer mündlichen Prüfung:

A framework for describing the speaking construct

Language competence
Phonology: Pronunciation, Stress, Intonation
Accuracy: Syntax, Vocabulary, Cohesion
Fluency: Hesitation, Repetition, Re-selecting inappropriate words, Re-structuring sentences, Cohesion
Strategic capacity
Achievement strategies: Overgeneralization, Paraphrase, Word coinage, Restructuring, Cooperative strategies, Code switching, Non-linguistic strategies
Avoidance strategies: Formal avoidance, Functional avoidance
Textual knowledge
The structure of talk: Turn taking, Adjacency pairs, Openings and closings
Pragmatic knowledge
Appropriacy; Implicature; Expressing being
Sociolinguistic knowledge
Situational; Topical; Cultural
(nach Fulcher 2003: 48)

Nach Kriterien der Ausführlichkeit ist eine solche Aufstellung kaum zu überbieten; Sachzwänge der Handhabbarkeit dürften wohl bei der Einschätzung der Brauchbarkeit einer solchen detaillierten Aufstellung allerdings wohl zum Wunsch nach einem etwas vereinfachten Entwurf Anlass geben. Doch gerade die angestrebte Klarheit, die den Ansatz Fulchers hier charakterisiert, kennzeichnet auch seine konkreten Vorschläge zur Beschreibung von Aufgaben, die seiner Meinung nach für mündliche Prüfungen geeignet sind:

A framework for describing tasks

1. Task orientation
- Open: outcomes dependent upon speakers
- Guided: outcomes are guided by the rubrics, but there is a degree of flexibility in how the test taker reacts to the input.
- Closed: outcomes dictated by input or rubrics

2. Interactional relationship
- Non-interactional
- Interactional:
  - One-way
  - Two-way
  - Multi-way

3. Goal orientation
- None
- Convergent
- Divergent

4. Interlocutor status and familiarity
- No interlocutor
- Higher status
- Lower status
- Same status
- Degree of familiarity with interlocutor

5. Topic(s)

6. Situations

(Fulcher 2003: 57)

Die Frage nach einem geeigneten Konstrukt für das Prüfen und Bewerten mündlicher Leistungen ist von zentraler Bedeutung bei Entscheidungen zur Aufgabenstellung. Diese Entscheidungen bestimmen die Prüfungsdurchführung mit, wie wir im nächsten Abschnitt sehen werden. Die Entwicklung eines passenden Konstrukts ist nicht zuletzt deswegen von solcher Wichtigkeit, als das Beschreiben typischer Merkmale mündlicher Interaktion entscheidende Auskunft darüber erteilt, in welchem Maße beziehungsweise ob überhaupt Bewertungsinstrumente wie Deskriptoren, die sich bei schriftlichen Leistungen bewährt haben, eine Anwendung im mündlichen Bereich finden können.

1.4 Prüfungsdurchführung

> „[T]he interviewer typically speaks far more than the test taker“
>
> (Alderson / Banerjee 2002: 93, Ergebnisse von Merrylees / McDowell (1999) referierend)

Manche Aspekte der Durchführung mündlicher Prüfungen werden bereits durch die jeweilige Aufgabenstellung vorgegeben; andere wiederum, wie die Wahl des Gesprächspartners oder der Gesprächspartner, müssen extra bedacht werden. Die Wichtigkeit des bedachten Planens der Prüfungsdurchführung beruht darauf, dass, wie die Wissenschaft immer wieder empirisch nachweist, die Wahl für die eine oder die andere Durchführungsmöglichkeit nicht weniger als das Entscheiden für die eine oder andere Art von Aufgabenstellung durchaus individuell recht unterschiedliche Auswirkungen auf das potentielle Verhalten und dadurch auch die erreichbare Leistungserbringung eines Prüflings haben kann. Das einleitende Zitat dieses Abschnitts von Alderson und Banerjee weist bereits auf die sehr bekannte, wenn auch nicht immer bewusste Neigung von Lehrern und Prüfern hin, Gespräche durch Verlängerung der eigenen Redeanteile zu dominieren. Auf Grund solcher beobachtbarer Verhaltensweisen ist man als Testanbieter gehalten, sich gut zu überlegen, welche Konstellation von Gesprächsteilnehmern für die jeweilige Prüfung am geeignetsten ist.

Wenn man verschiedene mündliche Prüfungen vergleicht, die von unterschiedlichen Anbietern gehalten werden, stellt man fest, dass es eine Reihe von Möglichkeiten gibt. Eine Möglichkeit sind Prüfungsgespräche zwischen einem oder

mehreren Prüflingen und einem oder mehreren Prüfern. Eine andere Möglichkeit ist, das Gespräch zwischen Prüflingen stattfinden zu lassen, wobei in diesem Fall die Prüfer als Beobachter fungieren. Es kommen auch Mischformen vor, in denen etwa ein einleitendes Gespräch zwischen Prüfer und Prüflingen geführt wird, bevor die Prüflinge versuchen, gemeinsam eine mündliche Aufgabe zu lösen. Unterschiedlich ist auch – teils aus organisatorischen Gründen – die Entscheidung, wie viele Prüflinge gleichzeitig geprüft werden sollen. Bei der Wahl der geeignetsten Form für die jeweilige Stufe einer KMK-Zertifikatsprüfung sollen neben den Erfahrungswerten der Praxis auch die Ergebnisse wissenschaftlicher Untersuchungen berücksichtigt werden. Alderson und Banerjee führen etwa in dem oben einleitend zitierte Übersichtsaufsatz fort:

> [T]here are concerns about the paired format, particularly with respect to the relationship between test takers (Foot, 1999) and the effect of test taker characteristics on test performance […]. This […] suggests that it may not be fair to assign scores to individuals in group assessment
> (Alderson / Banerjee 2002: 94; die Autoren nehmen unter anderem auch auf die Arbeiten von Morton 1998 und Swain 2001 Bezug).

Hier führt die Theoretisierung sowohl der Frage des Verhältnisses zwischen Testteilnehmern als auch der Persönlichkeitsmerkmale individueller Prüflinge zur konkreten Infragestellung der Validität der Benotungspraxis bei mündlichen Gruppenprüfungen. Aus unterrichtspraktischen Motivationsforschungen, die er zusammen mit Kormos durchgeführt hat (Dörnyei / Kormos 2000), berichtet Dörnyei analog hierzu von

> a positive relationship between L2 learners’ willingness to engage in communicative tasks and (a) the speakers’ social status and (b) the quality of the social relationship between the speaker and the interlocutor
> (Dörnyei 2001: 81).

Aber wie lassen sich von solchen weit reichenden theoretischen Analysen konkrete Empfehlungen für die Prüfungspraxis ableiten? Ist man nicht leicht entmutigt von dem Umfang potentieller Berücksichtigungsfaktoren, wie man auch im vorhergehenden Abschnitt von der Detailliertheit des Fulcher’schen Konstruktrahmens nur allzu schnell eingeschüchtert werden kann?

Es gibt zwar ausdrückliche Praxisempfehlungen in der Literatur, doch nicht alle werden wohl zwangsläufig den Beifall aller Prüfungsdurchführenden genießen dürfen. Nachdem Norton feststellt, dass die Frage, wie man sich entscheidet, welche Prüflinge man bei Paarprüfungen zusammen prüfen soll, noch „more detailed consideration and further research“ verdiene (Norton 2005: 287) plädiert sie dafür, bei der Wahl, ob Paar- oder Einzelprüfungen stattfinden sollen, bei den Prüflingen selbst Rat zu holen:

> Following Együd and Glover (2001), it would seem worthwhile to find out student preferences for paired or individual interviews in speaking tests
> (Norton 2005: 295).

Doch andere Fachempfehlungen werden bei Prüfern vermutlich auf mehr Verständnis treffen: Wie Lee im vorherigen Abschnitt zur Anzahl der zu stellenden Aufgaben einen konkreten Vorschlag unterbreitet, so stellt Tscherner zur empfehlungswerten Zahl der Prüfer fest,

> dass sich die Zuverlässigkeit der Bewertung deutlich erhöht, je mehr Bewerter dieselbe Prüfung beurteilen […] Kenyon / Tschirner (2000) ziehen daraus den Schluss, dass bei mündlichen Prüfungen mindestens zwei Bewerter unabhängig voneinander ihre Bewertungen abgeben müssen (Tschirner 2001: 109-110).

Nun erscheint es gegenwärtig aber einfach so zu sein, dass es leichter ist, auf empirischer Grundlage eine Empfehlung für die Anzahl der Prüfungsaufgaben und Prüfer zu geben als Konsequenzen zu ziehen aus der Tatsache, dass etwa „personality factors" (Fulcher 2003: 49) oder „variability in interlocutor style" (Brown 2003, 2004, zitiert nach Lumley / O'Sullivan 2005: 433) einen Einfluss auf die Entwicklung eines Prüfungsgesprächs haben (siehe jetzt auch Brown 2005). Zu den Rollen von Gesprächsteilnehmern stellt Wigglesworth fest: „the role of the interlocutor […] is *central* in ensuring that learners obtain similar output across similar tasks" (Wigglesworth 2001: 189). Damit ist aber noch gar nichts darüber gesagt, wer ein „interlocutor" sein darf und soll. ALTE members definieren diesen Begriff allein aus der Optik des Testens:

> In a test of speaking, the examiner who explains the tasks, asks questions and generally interacts orally with the candidate(s)
> Alte members 1998: 148.

Als deutsche Übersetzung hierfür wird „Fragesteller" empfohlen (ebd.: 125).

Die zweite Definition, die Davies et al. anbieten, bezieht sich ebenfalls direkt auf mündliche Prüfungen:

> 2. In language testing the term is used to refer to the interviewer or facilitator of communication in an oral interview (Davies et al. 1999: 85).

Die erste Definition an dieser Stelle führt jedoch eine Möglichkeit auf, die bei ALTE members fehlt. Diese Variante spiegelt nicht nur den Alltagsgebrauch des Ausdrucks wider, sie stellt ebenfalls eine verbreitete alternative Prüfungsvariante dar:

> 1. Any person who makes an active spoken contribution to a conversation or some other form of oral interaction (ebd.).

Wie auch immer die Entscheidung bei verschiedenen Testanbietern fällt, Prüfer als Gesprächsteilnehmer zuzulassen oder die mündliche Interaktion unter Prüflingen stattfinden zu lassen, wichtig dürfte es für Prüfer etwa sein, sich bei der Prüfungsdurchführung wie bei der anschließenden Beurteilung bewusst zu sein, dass sowohl Prüfer, sofern bei der Prüfung aktiv beteiligt, als auch verschiedene Prüflinge durch ihre Persönlichkeitsmerkmale und ihr individuelles Verhalten einen Einfluss auf den Ablauf der Prüfung haben. Dies dürfte vor allem beim ersten Fall von besonderer Bedeutung sein, wenn Prüfer selbst aktiv am Prüfungsgespräch beteiligt sind, denn „teacher talk" ist nicht nur ein bekanntes und natürlich auch nötiges Phänomen des Klassenzimmers; auch in der Prüfungssituation verhält sich mancher Prüfer wohlwollend und entgegenkommend in einer Art und Weise, die nicht zwangsläufig einer beruflichen Wirklichkeit entspricht. Die Ergebnisse entsprechender Studien fasst Fulcher wie folgt zusammen: „interlocutor raters accommodate their speech to that of the test taker" (Fulcher 2003: 49). Damit soll keine Kritik an dieser Prüfungshaltung ausgesprochen worden sein. Wichtig ist aber wohl, dass sich Prüfer, die sich am Prüfungsgespräch beteiligen, bewusst sind, in welcher Art und Weise sie dies tun und welche Auswirkungen dies auf die angestrebte Authentizität des Berufsbezugs der Prüfung haben kann. Damit soll nicht gesagt worden sein, dass Prüflinge nicht selbst die Fähigkeit zeigen können, ihrem beispielsweise sprachlich schwächeren Gesprächspartner gegenüber Entgegenkommen zu zeigen; doch die täglichen Lehrgewohnheiten fließen vermutlich stärker ins Prüfungsgespräch eines Lehrers ein als die auch aus beruflicher Sicht zu begrüßende Sensibilität mancher Prüflinge. Diesem Unterschied soll man sich als Prüfer bewusst werden; bei der Bewertung allemal, aber auch bereits bei der Entscheidung für die geeignete Form der Prüfungsdurchführung.

1.5 Bewertung

Wir haben uns oben einzelne Überlegungen zum Entwerfen von geeigneten Testaufgaben für mündliche Prüfungen angesehen sowie uns darüber Gedanken gemacht, welche Aspekte bei der Durchführung solcher Prüfungen zu berücksichtigen sind. Im jetzigen Abschnitt beschäftigen wir uns nun mit dem Teil des Prüfungsgeschehens, bei dem wohl am meisten Tinte fließt: die Bewertung.

Der Grund, warum so viel über das Bewerten geschrieben wird, zudem von mündlichen Prüfungen, ist schlicht darin zu finden, dass dieses Bewerten so schwierig ist. Bygate et al. sprechen in diesem Zusammenhang des „assessment of spoken language performance“ denn auch unverblümt von „the difficulty it entails“ (Bygate et al. 2001: 163).

Neben dem oben zitierten allgemeinen Rahmen eines Testkonstrukts für mündliche Prüfungen, den Fulcher 2003 vorlegte, wurde auch ausdrücklich für den berufsbildenden Schulbereich der Schweiz ein „Kriterienraster zur Beurteilung mündlicher Leistungen im Fremdsprachenunterricht“ erarbeitet (Ghisla / Kolb 2001:60). Die Autoren dieses Rasters drücken die Problematik in fast dem gleichen Wortlaut wie Bygate et al. aus: „Erfahrungsgemäss erweist sich die Beurteilung mündlicher Leistungen als schwierig“ (ebd.).

Da im dritten Jahr des Modellversuchs sehr viel Zeit darauf verwendet wurde, verschiedene Deskriptoren zur Bewertung von mündlichen Prüfungen zu vergleichen und gemeinsam weiter zu entwickeln möchte ich an dieser Stelle einleitend die Arbeit der Schweizer Kolleginnen und Kollegen in diesem Bereich dokumentieren. Es handelt sich um ein Raster zur Beurteilung mündlicher Leistungen zum Erwerb der Berufsmaturität (man beachte, dass in der Schweiz die Note 6 die beste ist):

Zur Beurteilung mündlicher Leistungen (Beispiel)

Grammatik **Die Äusserungen ...**

Note 6 sind weitgehend fehlerfrei.
Note 5 enthalten einige Fehler, die das Verstehen aber nicht beeinträchtigen.
Note 4 enthalten mehrere Fehler, die das Verstehen etwas beeinträchtigen.
Note 3 enthalten Fehler, die das Verstehen erheblich beeinträchtigen.
Note 2 sind weitestgehend unverständlich.

Wortschatz **Der Wortschatz ist ...**

Note 6 variationsreich; fehlende Begriffe können problemlos umschrieben werden.
Note 5 der Aufgabe angemessen; fehlende Begriffe können in den meisten Fällen umschrieben werden.
Note 4 angemessen, aber einfach; fehlende Begriffe können annähernd umschrieben werden.
Note 3 in mehreren Fällen der Aufgabe zu wenig angemessen.
Note 2 für die Aufgabe nicht ausreichend.

Aussprache und Intonation **Aussprache und Intonation weisen ...**

Note 6 keine wesentlichen Abweichungen von gesprochener Standardsprache auf.
Note 5 einige Abweichungen auf, die das Verstehen jedoch nicht beeinträchtigen.
Note 4 Abweichungen auf, die das Verstehen gelegentlich beeinträchtigen.
Note 3 Abweichungen auf, die das Verstehen häufig beeinträchtigen.
Note 2 so starke Abweichungen auf, dass das Verstehen stellenweise unmöglich ist.

Interaktives Verhalten **Die/Der Lernende hält das Gespräch ...**

Note 6 selbst in Gang.
Note 5 weitgehend selbst in Gang.
Note 4 mit einiger Hilfe in Gang.
Note 3 nur dank wiederholter Hilfe einigermassen in Gang.
Note 2 Es kommt kein richtiges Gespräch zu Stande.

Inhalt/Aussage **Die Ausführungen sind inhaltlich ...**

Note 6 treffend, ausführlich und durchwegs dem erforderlichen Niveau entsprechend.
Note 5 angemessen und ausführlich, dem Niveau entsprechend.
Note 4 meistens angemessen und dem Niveau einigermassen entsprechend.
Note 3 oft unangemessen bzw. zu knapp.
Note 2 meistens zu wenig relevant.

(Ghisla / Kolb ebd.)

Sowohl Fulcher als auch Ghisla und Kolb verwenden fünf verschiedene Kategorien, die berücksichtigt werden sollen. Manche, die sich etwa wegen der Handhabbarkeit des Bewertungskriteriums eine Einschränkung solcher Kategorien wünschen, mögen sich bei den Empfehlungen Skehans wohler fühlen, der sich auf drei Bereiche konzentriert: „measures are required in the three areas of complexity, accuracy and fluency" (Skehan 2001: 170). Wer sich bereits darüber freut, während einer Prüfung nur drei statt fünf Kategorien gleichzeitig berücksichtigen zu müssen, muss nun weiter lesen und beherzigen, dass es Skehan in seinen Empfehlungen nicht um eine Beurteilung einzelner Fertigkeiten unabhängig von einander geht, sondern dass auch das Zusammenspiel aller drei Bereiche mitberücksichtigt werden muss: „the competition between them will have an important impact on decisions that are made" (ebd.). (Skehans Ausführungen beziehen sich an dieser Stelle sowohl auf die Aufgabenstellung als auf die Bewertung.)

Bei Empfehlungen dieser Art wünscht man sich als Prüfer konkrete Handlungsanweisungen. Skehan versucht diesem Wunsch zu entsprechen indem er in einer Fußnote genauer erläutet, was er unter dem Begriff „complexity" versteht:

> The construct of complexity is close to what testers mean by range, in that both focus on a willingness to use a greater variety of syntactic forms
> (ebd.: 183).

Die Äußerungen Shehans beziehen sich hier auf die Bewertung jeder Art fremdsprachlicher Leistung und nicht speziell von mündlichen Leistungen. Luchtenberg hingegen versucht, unsere Aufmerksamkeit auf die für diesen dritten Teil der Handreichung wichtige Unterscheidung zwischen schriftlicher und mündlicher Kommunikation zu lenken in dem sie selbst den etwas heiklen Begriff der Sprachnorm problematisiert:

> Das wesentliche Kriterium für die Bewertung von Sprachäußerungen ist die Sprachnorm. Diese klare Aussage verliert allerdings an Eindeutigkeit, wenn die Unterschiede zwischen gesprochener und geschriebener Sprache stärker berücksichtigt werden, denn die ‚Sprachrichtigkeit' in mündlicher Rede verstößt auch bei Muttersprachlern in der Regel gegen die Norm der geschriebenen Sprache
> (Luchtenberg 2002: 85).

Luchtenbergs Äußerungen sind charakteristisch für eine ganze Reihe von wissenschaftlichen Beiträgen, die immer wieder betonen, dass wir es mit der Bewertung von mündlichen Leistungen mit einer ganz anderen Art der Kommunikation zu tun haben. Die Folge dessen ist, dass unsere Bewertungsinstrumente explizit auf die Eigenschaften natürlich vorkommender mündlicher Äußerungen basie-

ren müssen und keineswegs eine einfache Übertragung bewahrter Bewertungsinstrumente aus dem Bereich der schriftlichen Prüfungen zulässig ist.

Hughes beschreibt ausführlich, wie wenig mündliche Interaktion oft den Kriterien entspricht, die man üblicherweise auf die Bewertung der Qualität schriftlicher Leistungen anwendet:

> Spontaneous interactive speech will be full of hesitations, false-starts, grammatical inaccuracies, have a limited vocabulary, tend towards repetition and be structured around short thought units or quasi-clauses based on the constraints of breath and of spoken language processing
> (Hughes 2002: 77; Vergleiche auch Weir 2005: 107).

Eine empirische Untersuchung, die zeigt, wie unterschiedlich das Bewertungsverhalten verschiedener Gruppen sein kann, wurde 2003 von Sundh vorgelegt. Dieser fand heraus, dass Muttersprachler des Englischen bei der Bewertung mündlicher Leistungen mehr Wert auf kommunikative Fertigkeiten legten, während für schwedische Lehrer grammatikalische Korrektheit wichtig war:

> The Native Speakers focused on the students’ communicative skills and the messages conveyed, while the School Teachers […] tended to pay more attention to grammatical accuracy and the students’ grammatical repertoire (Sundh 2003: 265).

(Im Abschnitt II. 1.4 oben sahen wir bereits, dass Sundh herausfand, dass die Muttersprachler in seiner Untersuchung insgesamt weniger streng benoteten als die schwedischen Lehrer.)

Aus dieser Übersicht soll ersichtlich sein, dass die Wissenschaft sich wiederholt bemüht, sowohl umfangreiche Kriterienkataloge für die Bewertung spezifisch mündlicher Sprachleistungen zu erstellen als auch darauf hinzuweisen, dass die Gefahr besteht, ohne Berechtigung in der muttersprachlichen Sprachpraxis die Bedeutung des einen oder anderen Bereichs künstlich höher anzusiedeln. Stichfeste Empfehlungen, wie vom Prüfer gewährleistet werden soll, dass das anzustrebende Gleichgewicht zwischen den verschiedenen Bewertungskategorien erreicht werden soll, sind das noch lange nicht. Wenn etwa Fulcher in seinem bei III. 1.3 oben zitierten Konstrukt von „pragmatic knowledge“ und „sociolinguistic knowledge“ spricht, wie sind diese zu definieren, von einander zu unterscheiden und zu beurteilen? Im Bereich der „inter-language pragmatics“ spricht Hughes mit Bezug auf Rose und Kasper (2001) von „issues surrounding differences in conversational behaviour““ (Hughes 2002: 49). Doch wie sollten diese in die Bewertung mit einfließen? Und nach welchen „Normen“ würde man bei einer KMK-Zertifikatsprüfung urteilen? Nach den englischen? Den amerikanischen? Den deutschen? Den türkischen? Wer nimmt an der Prüfung teil und welche Zielgruppe gilt als Modell?

Sich auf Jones (2001) beziehend vertritt Luoma folgende These:

> Good storytelling routines are important for speakers, as one of the most common types of chatting involves personal stories about accidents or embarrassing situations (Luoma 2004: 24).

Doch ist diese Fähigkeit in jedem Kulturkreis gleich ausgeprägt? Wie wertet man als deutscher KMK-Prüfer die rhetorischen Schreibstrategien des irischen Verfassers dieses Handreichungsteiles? Oder machen wir uns vielleicht allzu viele Sorgen um die Treffsicherheit unserer Kategorien, wenn eventuell alles tatsächlich viel enger zusammenhängen sollte als bisher oft angenommen. Beebe und Zhang Waring behaupten auf jeden Fall auf Grund ihrer empirischen Untersuchungen, dass die Bereiche Pragmatik und Grammatik eng miteinander verbunden sind: „grammatical proficiency does contribute to pragmatic proficiency“ (Beebe / Zhang Waring 2004: 243).

Was zeigen uns diese Beispiele aus der Fachliteratur zum Thema Bewertung mündlicher Prüfungen?

Dass es sich hier wohl um einen noch umstritteneren Bereich als den der Bewertung schriftlicher Leistungen handelt; und dass die Bemühungen aller Beteiligten im dritten Jahr des Modellversuchs EU-KonZert zur Sicherung der Vergleichbarkeit der Standards zu gemeinsamen Ergebnissen zu kommen, ein nicht weniger mutiges als auch nötiges Unterfangen war.

1.6 "Oh, and another thing I wanted to say was…"

Wie am Anfang dieses Handreichungsteiles versprochen, komme ich jetzt auf die bei III. 1.1 abgedruckten Transkripte von Prüfungsauszügen zu sprechen. Aber warum wohl, wenn ich schon Abschnitte zu den drei Aufgaben dieser Modellversuchsphase „Aufgabenstellung", „Prüfungsdurchführung" und „Bewertung" geschrieben habe?

Jeder Prozess der gemeinsamen Standardfindung muss sich um die Validierung der gemeinsam erzielten Ergebnisse bemühen. Im Falle des KMK-Fremdsprachenzertifikats handelt es sich zudem um Prüfungen, die zu den ersten und erst recht zu den ersten berufsbezogenen Prüfungen in Deutschland gehören, die sich an die Empfehlungen des Gemeinsamen europäischen Referenzrahmens angelehnt haben. Verständlicherweise und in bester bildungspolitischer Absicht regt der Europarat jetzt an, dass überall dort, wo mit dem Stempel Gemeinsamer europäischer Referenzrahmen / Common European Framework / Cadre européen commun hantiert wird, vom Rahmen viel enthalten ist. Es geht bei den parallel laufenden Vorhaben „Weiterentwicklung des Referenzrahmens" und „Sicherung der Vergleichbarkeit der Standards der KMK-Fremdsprachenzertifikatsprüfungen" um das gemeinsame Ziel der Validierung der jeweiligen Prüfungsentwürfe. Bei der Validierung einer Prüfung geht es um den „Prozess der Sammlung von Belegen für die Richtigkeit von aus den Testwerten gezogenen Schlussfolgerungen" (ALTE members 1998: 130). Auch der Europarat betont die Notwendigkeit, Folgendes zu tun: „outlining the validation process and providing evaluation criteria for the technical quality of the standard setting" (Takala 2004: iv). In "Section A: Overview of the Linking Process" des gleichen *Reference Supplement* zum *Manual for Relating Language examinations to the Common European Framework* wird das Stichwort „Empirical Validation" beschrieben als

> the collection and analysis of test data and ratings from assessments to provide evidence that both the examination itself and the linking to the CEFR are sound" (Takala (ed.) 2004: 1).

Bei diesem *Reference Supplement* des Europarates ist bisher nur bei der Nachweismöglichkeit der "Specification" eine qualitative Vorgehensweise zur Validierung einer Prüfung vorgesehen:

> To prove internal and external validity, quantitative and qualitative methods can be combined. **Specification** (Chapter 4) can be seen as a qualitative method: providing evidence through content-based arguments (ebd.: 2).

Wenn dort von empirischer Validierung gesprochen wird, ist hingegen nur von quantitativen Prozeduren die Rede (ebd.; als Beispiele für Daten, die zur quantitativen Analyse der Nützlichkeit eines Tests herangezogen werden können führt etwa Bachman ‚test scores, scores for items or tasks, or responses to questionnaires and self-ratings' auf [Bachman 2004: 6]). Wenn das *Reference Supplement* des Europarates von empirischer Validierung spricht kommt allerdings leider nicht zur Geltung, dass neuere Forschungsarbeiten dafür plädieren, zur Ergänzung von tra-

ditionellen statistischen Methoden auch von anderen, qualitativen, Forschungsansätzen Gebrauch zu machen:

> Language testers have generally come to recognize the limitations of traditional statistical methods for validating oral language tests and have begun to consider more innovative approaches to test validation, approaches that promise to illuminate the assessment process itself, rather than just assessment outcomes (i.e., ratings) (Lazaraton 2002: xi).

Lazaraton selbst empfiehlt einen solchen möglichen qualitativen Forschungsansatz bei der Validierung: „One such approach is conversation analysis (CA)“ (ebd.). Diese Forderung rechtfertigt sie mit folgender Feststellung:

> Conversation analysis provides a uniquely suited vehicle for understanding the interaction, and the discourse produced, in face-to-face oral assessment procedures (ebd.: 171).

Ellis und Barkhuizen erklären den Ansatz der Konversationsanalyse in einer etwas vorsichtigen Art und Weise, die dennoch seine Bedeutung für die Bewertung von mündlichen Sprachleistungen klar verdeutlicht, wie folgt:

> Ordinary conversation […] is what other types of talk are measured against, since, CA purists would argue, it is ‘the predominant form of human interaction in the social world and the primary medium of communication to which the child is exposed and through which socialization proceeds’ (Heritage 2001: 2741).
> (Ellis / Barkhuizen 2005: 200).

Dabei heben Ellis und Barkhuiszen hervor, dass dieses Forschungsparadigma der Konversationsanalyse von Anfang an trotz der durch Textverarbeitungstechniken möglich gewordenen elektronischen Analyse von Verbaldaten von quantitativen Untersuchungsansätzen Abstand gehalten hat:

> It makes sense, therefore, that the quantification of the data, the counting of patterns and structures, does not have a place in the analysis, as Heritage (2001: 2744) remarks, ‘statistical analysis has played little role in the field, largely because in the matter of interactional practices, as in the case of biological species, large numbers are not essential to establishing their existence’
> (Ellis / Barkhuizen 2005: 212).

Wie könnte eine konversationsanalytische Untersuchung der drei Prüfungsausschnitte bei 1.1 aussehen und welcher Bewertungsschemata würde man sich bedienen, um eine solche Untersuchung zu unterstützen?

Da ein großer Teil der Arbeit der dritten Phase des Modellversuchs darin bestand, verschiedene Deskriptorensätze zu erproben und weiterzuentwickeln, ohne dass abschließend eine Empfehlung für den einen oder anderen Satz als verbindlicher ausgesprochen wurde, werde ich mich in meinen kurzen verbleibenden Bemerkungen zu den PrüfungsTranskripten auf die Hinweise des Gemeinsamen europäischen Referenzrahmens beziehen, der den Ausgangspunkt der Bemühungen der Zertifikatsprüfungen bildet. Aus Platzgründen erheben die folgenden Ausführungen keineswegs den Anspruch, eine vollständige konversationsanalytische Behandlung der bei 1.1 abgedruckten Transkripte zu sein. Es geht in diesem abschließenden Abschnitt lediglich darum, anzudeuten, erstens, wie schwer die Bewertung von mündlichen Sprachleistungen überhaupt ist und zweitens, welche Beiträge ausführliche konversationsanalytische Untersuchungen zur weiteren Validierung der KMK-Zertifikatsprüfungen noch leisten könnten.

Die erste bei 1.1 abgedruckte Aufgabe simuliert ein Telefongespräch in dem ein Kunde eine Lösung für sein Problem mit einer erworbenen CD-ROM sucht und

der Kundenberater dem Kunden erklären will, wie das Problem zu beheben ist. Wie bewältigen die zwei Prüflinge die Aufgabe nach der Kategorie „Spektrum“ („range“), der ersten der fünf Kategorien des Beurteilungsrasters des Referenzrahmens zur mündlichen Kommunikation „Spektrum“ („range“) (Trim et al. 2001: 37, 38)?

Beim Spektrum heißt es im Referenzrahmen:

A2: Verwendet elementare Satzstrukturen mit memorierten Wendungen, kurzen Wortgruppen und Redeformeln, um damit in einfachen Alltagssituationen begrenzt Informationen auszutauschen.

B1: Verfügt über genügend sprachliche Mittel um zurechtzukommen; der Wortschatz reicht aus, um sich, wenn auch manchmal zögernd und mit Hilfe von Umschreibungen, über Themen wie Familie, Hobbys und Interessen, Arbeit, reisen und aktuelle Ereignisse äußern zu können.

B2: Verfügt über ein ausreichend breites Spektrum von Redemitteln, um in klaren Beschreibungen oder Berichten über die meisten Themen allgemeiner Art zu sprechen und eigene Standpunkte auszudrücken; sucht nicht auffällig nach Worten und verwendet einige komplexe Satzstrukturen.
(nach Trim et al. 2001: 37, 38).

Der Kunde B. beschreibt sein Problem wie folgt:

Em, yes.
Eh, I put the CD Rom X 4 hundred 7 like the, the
[iks]
[EXHALES LOUDLY]
directory in my manual in the computer,
and tried to INstall the program on the disk which was in the car-, cartoon
eh where the CD Rom was
and then cames the arrow, error
[THREE-SECOND PAUSE]
CD Rom not found.
I don’t know what I shall do?

Der Kundenberater A. schlägt eine Lösung in folgendem Wortlaut vor:

Oh, good,
em,
you go in the internet on the site tribble w dot western tech dot com.
And you go on the link to downloads.
You download the file bug fix d u 4 point 1 and after these you have downloaded
the file you installed these program
After the installation you em start the computer at new and then em
the problem don’t, don’t were on your computer.

B.’s “I […] tried to Install the program on the disk which was in the car-, cartoon eh where the CD ROM was and then cames the arrow” klingt beim ersten Hinhören schon recht elementar und ebenso A.’s “after the installation you em start the computer at new and then em the problem don’t, don’t were on your computer” ebenso. Also sind beide auf Stufe A2 des Referenzrahmens oder Stufe I der KMK-Prüfung?

Bei B1 soll man hingegen sprachlich zurecht kommen. Aber was heißt es, sprachlich zurecht zu kommen? Der Kunde beschreibt sein Problem mittels Sprache. Die Sprache ist zwar fehlerhaft, aber es gelingt dem Kunden, sein Anliegen genau zu schildern. Das Gleiche gilt für die angebotene Lösung des Kundenberaters: es ist ein brauchbarer Vorschlag, in fehlerhafter Sprache erfolgreich ausgedrückt. Also sind beide Prüflinge auf der GER-Stufe B1 /KMK Stufe II?

Erlauben ihre Redemittel den Prüflingen, klare Beschreibungen zu formulieren? Falls ja, müsste man bei einer kriterienorientierten Prüfung nach dem Gesichtspunkt des verwendeten Sprachspektrums sogar das Niveau GER Stufe B2 / KMK Stufe III bescheinigen. Nun ja: Entscheidend ist, was man unter „klare Beschreibungen“ meint. Wie *klar* heißt „klar“? Welche Rekonstruktionsleistung des Zuhörers muss stattfinden, damit die intendierte Botschaft beim Empfänger wie beabsichtigt ankommt?

Wichtig ist auch, dass man die jeweilige Leistung tatsächlich nach den vorgegebenen Kriterien beurteilt. Habe ich in meiner Diskussion des verwendeten Sprachspektrums der beiden Prüflinge wirklich diesen Aspekt bewertet oder richtete ich mein Merkmal nicht bereits auf die zweite Kategorie des GERs zur Beurteilung der mündlichen Kommunikation, auf die Korrektheit? Der Umgang mit Deskriptorenkategorien will gelernt und geübt werden.

Beim zweiten Transkript versuche ich genau diese zweite Kategorie der Korrektheit zu beurteilen.

Der Referenzrahmen beschreibt mündliche Korrektheit wie folgt:

> **A2:** Verwendet einige einfache Strukturen korrekt, macht aber noch systematisch elementare Fehler.
>
> **B1:** Verwendet verhältnismäßig korrekt ein Repertoire gebräuchlicher Strukturen und Redeformen, die mit eher vorhersehbaren Situationen zusammenhängen.
>
> **B2:** Zeigt eine recht gute Beherrschung der Grammatik. Macht keine Fehler, die zu Missverständnissen führen, und kann die meisten eigenen Fehler selbst korrigieren (nach Trim et al 2001, 37-38).

Im zweiten Beispiel sollen die Prüflinge eine Reihe von Informationen miteinander austauschen, wobei C. die Rolle eines Praktikanten aus den Vereinten Staaten von Amerika übernimmt und D. die einer Angestellten der örtlichen deutschen Stadtverwaltung. Werden „einige einfache Strukturen korrekt“ verwendet (Stufe A2 / KMK I)? C. stellt sich korrekt vor: „My name is [C.]“. Er erklärt, dass er wegen eines Termins beim Bürgermeister am Nachmittag nur etwas später zu einer geplanten Besprechung mit der Angestellten kommen kann:

> Sorry, em, I can’t eh come to the meeting em before half past three.
> Em I have a me, I have em a meeting em with the mayor.

D. ihrerseits stellt sich ebenfalls korrekt vor, erläutert ihre Zuständigkeiten und begrüßt die ihr von C. angekündigte Unterstützung:

> Hello, my name’s [D.D.],
> I’m a council clerk from this administrative community.
> I’m responsible for the move and I’m VERY glad of your help.

Kann man bei beiden von einem „Repertoire gebräuchlicher Strukturen und Redeformen“ sprechen (Stufe B1 / KMK II)? Wie bei Transkript 1 oben, bei dem es darum ging, das verwendete sprachliche Spektrum zu beurteilen, sieht man hier beim zweiten Transkript, wie sehr der Prüfer bei der Anwendung jeder einzelner dieser Kategorien herausgefordert ist. Wie viele gebräuchliche Strukturen und Redeformen müssen denn dabei sein, um als „Repertoire“ zu gelten? Reichen hierzu solche Höflichkeitsbekunden C.’s wie „eh I’m here to help you“ oder D.’s ebenfalls höflich geäußerte Aufforderung „Can you bring some files to the financial administration for me please“?

Auf den ersten Blick scheint es leichter zu sein, die Frage nach der Zugehörigkeit zur GER-Stufe B2 / KMK-Stufe III zu beantworten: „Zeigt eine recht gute

Beherrschung der Grammatik" wenn C. an einer Stelle grammatikalisch fehlerhaft sagt „I can eh take this files em to the financial administration" oder D.'s ausführliche Wegbeschreibung ebenfalls von der Bahn der grammatikalischen Korrektheit abkommt:

> You must go upstairs to the first floor, then turn left and go till the end of the floor. There is the financial administration on your right hand side and next to mailroom.

Doch wie repräsentativ sind diese Beispiele für den jeweiligen Grad an Korrektheit, der den Gesamteindruck der Leistungen der Prüflinge Rechnung trägt? Wie umfangreich muss eine Prüfungsanalyse sein, welchen zeitlichen Umfang soll eine mündliche Prüfung haben, damit ein ausgewogenes Urteil gefällt werden kann? Gelingt es uns in unserer Bewertungspraxis zu erkennen, dass ein Prüfling im Laufe eines Prüfungsgesprächs vielleicht selbst registriert, welche grammatikalischen Fehler er gemacht hat (um diese später zu vermeiden), wenn wir uns als Prüfer nur das Vorkommen solchen Fehler notieren und nicht das spätere Fehlen hiervon im weiteren Verlauf eines Prüfungsgesprächs? Und wie viele von uns beurteilen die erbrachte Leistung während einer mündlichen Prüfung tatsächlich nach den Kriterien des Gebrauchs der mündlichen Sprache durch Muttersprachler, mit all den Verzögerungen, Wiederholungen, unvollendeten Formulierungen und gar grammatikalischen Fehlern, die, wie wir in unserer Literatursynopse oben gesehen haben, hierfür charakteristisch sein können? Oder vergleichen nicht viele von uns immer noch mündliche Fremdsprachenleistungen mit der in aller Ruhe ausgearbeiteten Prosa dieser Muttersprachler?

Als drittes Beispiel soll abschließend an Hand unseres dritten Transkripts die GER-Kategorie „Flüssigkeit" untersucht werden (Vergleiche hierzu die im Abschnitt 2.c1. beschriebenen Ergebnisse einer Diskussion dieser Aufnahme).

Bei diesem dritten Beispiel handelt es sich im Gegensatz zu den ersten beiden Transkripten um eine mündliche Prüfung mit drei Prüflingen. Die daraus erwachsende Herausforderung sowohl für Prüflinge als auch für Prüfer erkennt man bereits an der Komplexität der verwendeten Transkriptionskonventionen, die unter anderem deswegen so eigenartig wirken, weil es bei dieser dritten Aufnahme es anders als bei den beiden Zweiergesprächen so oft vorkam, dass Prüflinge gleichzeitig redeten.

(Dies wird im Transkript den Konventionen der Konversationsanalyse entsprechend durch Unterstreichen gleichzeitig gesprochener Redebeiträge signalisiert. Ferner werden solche simultan gesprochene Äußerungen bei der Durchnummerierung der Zeilen als eine Zeile nummeriert, obwohl durch die Gleichzeitigkeit des Sprechens räumlich zwei Zeilen benötigt werden, um diese Äußerungen zu transkribieren.) Bei der Kategorie „Flüssigkeit" wendet der Referenzrahmen folgende Deskriptoren an:

> **A2:** Kann sich in sehr kurzen Redebeiträgen verständlich machen, obwohl er/sie offensichtlich häufig stockt und neu ansetzen oder umformulieren muss.
>
> **B1:** Kann sich ohne viel Stocken verständlich ausdrücken, obwohl er/sie deutliche Pausen macht, um die Äußerungen grammatisch und in der Wortwahl zu planen oder zu korrigieren, vor allem, wenn er/sie länger frei spricht.
>
> **B2:** Kann in recht gleichmäßigem Tempo sprechen. Auch wenn er/sie eventuell zögert, um nach Strukturen oder Wörtern zu suchen, entstehen kaum auffällig lange Pausen
> (nach Trim et al. 2001: 37-38).

Zur Länge der Redebeiträge:

Bei Prüfling E. stellt man bei Zeilen 03-08 schon die Fähigkeit fest, länger zu sprechen (das heißt: besser als A2, also mindestens B1), ja sogar durchaus auch die, „in recht gleichmäßigem Tempo“ zu sprechen (B2; bei Transkript 3 fehlen, im Gegensatz zu den ersten beiden Transkripten, alle Hinweise auf Redepausen, die länger als eine Sekunde gedauert hätten, da solche Pausen bei diesem Ausschnitt gar nicht vorkamen):

03 E: eh maybe we eh have a little focus on Europe
04 So we have a lot of eh Europe Asian food, Europe mediterrean food,
05 but eh I think eh, my mind is coming now, eh
06 Polynesia,
07 never, never eh heard of eh Polynesian restaurants
08 so I don’t know what kind of food they have, so

Wenn damit Kandidat E. nachweist, auf KMK Stufe III zu sein, dann muss das wohl erst recht für Prüfling G. gelten:

27 G: just European?
28 I mean it’s a lot a lot of business in town, right, in the Europa Passage,
29 so a lot of offices and everything, so, em,
30 maybe we should do something which is kind of fast
31 and kind of eh good, fresh and everything
32 so whatever that the people come there for for lunch
33 then we really specialize on, on the lunch em time.
34 Like the Vapiano thing, which is already sold in pasta,
35 maybe we can do it with something else.
36 some, maybe Polonesisches, maybe one never know.
37 What-, whatever.

Bei Kandidat F. allerdings haben wir mit den vorliegenden Deskriptoren ein Problem, denn an keiner einzigen Stelle dieses Gesprächs weist F. die Fähigkeit nach, mehr als nur sehr kurze Redebeiträge leisten zu können. Doch wäre man berechtigt, Kandidat F. bei einer KMK-Prüfung der Stufe II oder III (GER-Stufe B1 und B2) durchfallen zu lassen? Sieht man sich Zeilen 14-20 an, so erkennt man, dass das Gespräch dieser Gruppenprüfung von E. und G. regelrecht dominiert wurde.

14 E: S-, something new. Never, never works
15 G: I mean downtown Hamburg here
F: OK
16 G: there’s a lot of business business
F: you can have it in the traders’ X
17 F: as well. Polynesian, Polynesian and European food
E: There’s Polynesian? Then let me take: New,
18 E: New Zealand? Oder Madagaskar
19 F: [EXHALES AIR]
20 E: there is the movie, the movie from the Walt Disney,
Madagaskar, with the yeah

In Zeile 15 zeigt sich F. als guter Zuhörer (“OK”). Nach G.’s abgeschlossener Feststellung aus den Zeilen 15 und 16 („I mean downtown Hamburg there’s a lot of business“) versucht F. dann selbst In den Zeilen 16 und 17, einen ausführlicheren Redebeitrag zu leisten: „you can have it in the traders’ X (Redebeitrag wegen der Unterbrechung G.’s an dieser Stelle unverständlich) as well. Polynesian, Polynesian and European food“. Doch F.’s Redebeitrag wird gleich zweimal unterbrochen. Zuerst signalisiert G., dass er F. doch nicht zu Wort kommen lassen will,

in dem er das Wort „business“ wiederholt. Nichtsdestotrotz versucht F. seinen Beitrag fortzusetzen, wird aber dann sehr dominierend von E. unterbrochen. In Folge dessen verzichtet F. bei Zeile 19 auf eine weitere Wortmeldung und gibt ein paraverbales Zeichen seiner Frustration von sich gibt („[EXHALES AIR]“). Diese Gelegenheit benutzt E., um mit seinem neuen Beitrag fortzufahren. Damit gelingt es E. und G. an dieser Stelle wieder, F. eine aktive, gestaltende Rolle in diesem Gespräch zu verwehren. Verdient er es wirklich, deshalb vom Verleihen eines Zertifikats der Stufe II oder III ausgeschlossen zu werden? Oder sind hier nicht andere Aspekte am Werk, die bei der Anwendung vorgesehener Deskriptoren berücksichtigt werden sollen?

Deskriptoren zur Beschreibung von Teilfertigkeiten können überarbeitet und verbessert werden. Doch eben so wichtig wie eine Überarbeitung scheint die Frage zu sein, wie die bewerteten Teilkompetenzen zueinander in Verbindung stehen. Entscheidend bleibt immer noch die Frage, welchen Stellenwert analytische und holistische Verfahren in unseren Bewertungen haben sollten. Diese Frage des Verhältnisses zwischen holistischen und analytischen Bewertungsverfahren wurde vor allem im Abschnitt II. 1.4 oben bereits ausführlich erläutert. Götz Reuter schildert in Teil c. der folgenden Betrachtung der Workshops des dritten Teiles des Modellversuchs (III. 2) die Erprobung und Weiterentwicklung der Bewertungsinstrumente, die in dieser letzten Modellversuchsphase stattgefunden haben.

In diesem Abschnitt ging es mir darum zu zeigen, dass das Entwickeln geeigneter Deskriptoren immer eine große Aufgabe ist; und ferner, dass die Anwendung solcher Deskriptoren von Prüfern geprobt und erlernt werden muss. Um dieses zu verdeutlichen, habe ich Prüfungstranskripte untersucht, ohne zu sagen, auf welcher Stufe diese Prüfungen stattgefunden haben. Dadurch konnte festgestellt werden, dass eine Zuordnung der Kandidaten zu den Stufen des europäischen Referenzrahmens beziehungsweise zu den Stufen der KMK-Prüfung nicht immer leicht war. Wie wir im Abschnitt II. 1.3 oben bereits gesehen haben, muss allerdings ein Prüfer immer daran denken, dass bei kriteriumsorientierten Tests „die Leistung eines Prüflings im Verhältnis zu einem zuvor definierten Kriterium interpretiert wird“ (ALTE members 1998: 110). Es bleibt einem Prüfling unbenommen, an einer Prüfung einer niedrigeren Stufe teilzunehmen, obwohl dieser Prüfling die besten Aussichten hätte, die Prüfung einer höheren Stufe zu bestehen. Bei kriterienorientierten Prüfungen kann jedoch nur die Erfüllung der jeweiligen Kriterien vom Prüfer bescheinigt werden, nicht jedoch das Vorhandensein von Fertigkeiten, die auf das Bestehen einer Prüfung einer höheren Stufe hoffen lassen.

Wie dem auch sei und für welche Stufe auch immer dieser freiwilligen Prüfungen unsere Prüflinge sich anmelden:

Wir wünschen uns alle bei unseren mündlichen Prüfungen Prüflinge, die nicht auf den Mund gefallen sind. Möge es uns gelingen, unsere Aufgaben so zu stellen und durchzuführen, dass wir bei unserer Bewertung möglichst oft begeistert ausrufen können: "Now you’re talking!"