Karin Vogt • Jürgen Quetz

Assessment

im Englischunterricht

Kompetenzorientiert beurteilen und bewerten

HELBLING

Innsbruck • Esslingen • Bern-Belp

Inhalt

Kapitel 1

Auftakt: Zur Kunst der Beurteilung sprachlicher Schülerleistungen

Im Titel dieses Buches wird das Wort „Assessment“ als übergeordneter Begriff für die unterschiedlichsten Dinge verwendet, die behandelt werden sollen. Assessment ist kein deutsches Wort, aber Sie kennen es sicher in Ausdrücken wie Assessment-Center. Wer in der Berufswelt ein Assessment-Center absolviert, nimmt an einem Testverfahren teil, in dem festgestellt wird, ob er oder sie für die Position geeignet ist, die eine Firma ausgeschrieben hat. Im Schulalltag ist es an den Lehrerinnen und Lehrern[1], Lerngruppen und deren Leistungen zu beurteilen und zu bewerten. Sowohl für die Schüler als auch für die Lehrkräfte spielt Assessment eine zentrale Rolle. Dies beginnt schon mit Beginn des Unterrichts.

1 Beurteilung und Bewertung

Ein ganz normaler Mittwochmorgen. Frau S. kommt zum Englischunterricht in ihre Klasse. Hausaufgabe war, Vokabeln zu lernen. Die Lehrerin fragt stichprobenartig des schriftlichen Tests die Vokabeln heute mündlich ab. Einzelne Schüler werden befragt, Frau S. notiert sich kurz eine Note in ihr Notenbuch: „Well done, Oliver, you've answered all my questions!“ – „Oh, Anne, the correct word is *Native Americans* and not *Indians*. You obviously didn't pay attention last week when we talked about that.“ Danach steht das Üben der *-ing form* auf dem Programm. Linus hat scheinbar noch Schwierigkeiten mit der Anwendung von Präpositionen. Frau S. nimmt sich vor, das noch einmal gemeinsam zu erörtern. Nun ist das Textverständnis dran, es folgt Gruppenarbeit zum Lehrwerk. Frau S. gibt den Schülern knapp Feedback zu den Ergebnissen der Gruppenarbeit. „Group A has found out a lot of interesting details. One thing: for the presentation, however, you should have more eye-contact with your audience.“ Eine Kollegin lässt gerade die Schüler untereinander Feedback zu Lernergebnissen geben, das möchte Frau S. auch mal ausprobieren. Doch nun sind noch einige Aufgaben im Workbook zu machen. Außerdem ist die nächste Klassenarbeit schon für die kommende Woche angesetzt.

Dieses aus dem Arbeitsalltag herausgegriffene Beispiel zeigt deutlich, wie häufig Lehrkräfte Rückmeldungen oder auch *Feedback* zu der Qualität von Schülerbeiträgen geben. Sie beurteilen, indem sie Schülerbeiträge loben oder aber inhaltlich bzw. formal korrigieren. Fehler lassen sich schwer ignorieren und jede Fehlerkorrektur ist implizit oder explizit ein Akt der Beurteilung. In jedem Fall aber können Schüler den Äußerungen ihrer Lehrkraft entnehmen, wie zufrieden diese mit den Antworten ist. Ähnliches gilt auch für schriftliche Äußerungen, die häufig nicht mit einer Note bewertet werden, sondern ähnlich wie die mündlichen Äußerungen explizit oder implizit beurteilt werden. Bei Hausaufgaben erfolgt ebenfalls oft nur ein schnelles Lob (oder ein Tadel) ohne formelle Bewertung.

Die Überprüfung, ob Vokabeln gelernt worden sind, liegt schon an der Grenze zu den formelleren Lernerfolgskontrollen. Hier erfolgt oftmals eine Benotung und somit eine *Bewertung*. Das wird bereits

[1] Um den Lesefluss nicht zu behindern, wird im Folgenden vornehmlich die maskuline Form von Schüler oder Lehrer verwendet, Schülerinnen und Lehrerinnen sind dabei selbstverständlich inkludiert. Wir bitten um Verständnis.

daran deutlich, dass das Notenbuch eine wichtige Rolle spielt: was dort nicht an Bewertungen, meist in Ziffernform, festgehalten ist, kann nicht rekonstruiert werden und spielt deshalb bei der Bewertung von Schülerleistungen langfristig keine Rolle. Terminologisch gilt es zwischen einer Bewertung und der eher informellen Beurteilung zu unterscheiden. Beurteilungen können sich natürlich auch in schriftlicher Form finden, etwa wenn Schüler bei einer Übungsarbeit stat[illegible]lote Kommentare erhalten, die über Stär[illegible]un[illegible]chw[illegible]chen Auskunft geben. Hier ein Beispie[illegible]ne[illegible]assenarbeit.

(1) Ganzheitliche Beurteilung

Couch Potatoes

Some couch potatos haven't got friends. Their didn't know what they can do with the time. So many couch potatos sit on the couch and watch TV or play computer or Playstation. But some children are playing on the computer for money. And for some children will it be a addiction. They come from the school and go on the computer or TV. A friend of me are a couch potato. All the day she sits on the couch and watch TV. She don[illegible] want to go out with me or her friends. But her [illegible] mother did nothing. Her mother means th[illegible] okay. Sometimes I am a couch potato wh[illegible] am ill. It's often the Sunday when I am don[illegible] what I can do and then I am turn[illegible]e [illegible] on watch all the soaps. But I didn't ma[illegible]often. I think all the couch potatos don't kn[illegible]w w[illegible] can do with the time so they ma[illegible]the TV on and watch TV.

A nice little essa[illegible]hara ! The reader learns from it[illegible] are no couch potato, but you under[illegible]nd [illegible]y others sometimes are. A good idea: p[illegible]s should be interested in what th[illegible]child[illegible]o!

S[illegible]io[illegible]r the n[illegible]t test: please study

- tenses, especiall[illegible]he simple present
[illegible]ations (d[illegible]don't, not: do'nt – doesn't)

Just generally well done!

(2) Analytische Beurteilung

Couch Potatoes

Some couch p[illegible]s have[illegible]nds.
Their didn't [illegible] they can do with the time. So many c[illegible] po[illegible]t on the couch and watch TV or play c[illegible]uter [illegible] or Playstation. But some [illegible] on the computer for money. [illegible] some c[illegible]ldren will it be a addiction. They co[illegible] the school and go on the computer or TV. A [illegible]end of me are a couch potato. All the day she sits on the couch and watch TV. She don't want to go out with me or her

pl. = potatoes / haven't
they don't / their

ξ games (expr.)
play (Gr) / ? (What do you mean?)
become / an (Gr) / word order
turn
mine is
watches
doesn't

friends. But her mother did nothing. Her mother means that is okay. Sometimes I am a couch potato when I am ill. It's often the Sunday when I ~~am~~ don't know what I can do and then I ~~am~~ turn the TV on and watch all the soaps. But I didn't make this often. I think all the couch potatos do'nt know what they can do with the time so they make the TV on and watch TV.

doesn't do anything
thinks (that this is)

Gr!

do
-oes / don't [illegible]ove)
(same as seco[illegible]nce) / turn

Too many [illegible]kes, Barbara !
I hope the next [illegible]'ll be better!

Eine ganzheitliche Beurteilung einer Schülerleistung erfordert von der Lehrkraft ein genaues Verständnis der Stärken und Schwächen einer sprachlichen Leistung und ist somit zeitaufwändiger und vielleicht auch schwieriger als der bloße Verweis auf einzelne Fehler. Andererseits ist sie sehr hilfreich für Schüler, weil konkrete Hinweise auf Fehlerschwerpunkte gegeben werden, auf die sie bei weiteren Übungen ihre Aufmerksamkeit richten sollten.

Ähnlich [illegible]ihren die meisten Lehrerinnen und Lehrer bei [illegible]lichen Beiträgen. Um die Kommunikation [illegible]t zu stören, machen sie sich in Gedanken [illegible]zu inhaltlichen oder typischen formalen Schwächen der Äußerung und thematisieren diese dann zu gegebenem Zeitpunkt.
Es sind aber alles noch Beurteilungen, die sich nicht unbedingt in einer Bewertung durch eine Note niederschlagen müssen.

2 Test? Assessment? Klassenarbeit? Prüfung? – Begriffe und ihre Geschichte

Wie kommt es eigentlich zu der Vielfalt an Begriffen für im Grunde gleiche oder ähnliche Dinge? Zu Beginn des vorigen Jahrhunderts gab es in Deutschland eigentlich nur zwei Begriffe: die „Klassenarbeit" und die „(Abschluss-/Aufnahme-…)-Prüfung". Es folgt daher ein kurzer Abriss zur Entwicklung im Bereich von bzw. Assessments, der sich [illegible]setzt, Schneisen in den terminologischen [illegible]schlagen.

Traditionelle Assessments

In den Jahren vor dem 2. Weltkrieg war man der Meinung, dass sich die Beherrschung einer Sprache vor allem an der korrekten Lösung von Grammatikaufgaben feststellen lasse. Wirkliches Sprachkönnen unterstellte man aber erst, wenn die Lernenden in der Lage waren, Übersetzungen in die Zielsprache vorzunehmen oder Aufsätze zu schreiben. Diese traditionsreichen Prüfungsformen waren allerdings schon vor dem 2. Weltkrieg als höchst subjektive Formen der Leistungsmessung in die Kritik geraten. Aufsätze kann man bekanntlich kaum objektiv beurteilen, wie zahlreiche Studien zeigen (vgl. u.a. Ingenkamp 1971).

Noten für den gleichen Aufsatz oder die gleiche Mathearbeit, das hat man immer wieder festgestellt, schwanken bei verschiedenen Korrektoren fast über die ganze Notenskala, und Schüler können sich nicht sicher sein, dass die gleiche Leistung von der gleichen Lehrkraft zu verschiedenen Zeitpunkten auch gleich beurteilt wird. Schon die Position eines Klassenarbeitshefts innerhalb eines Stapels von 30 anderen Heften kann die Note beeinflussen, wie man weiß: Wir ermüden bei der Korrektur, verlieren den Überblick und werden, je nach Naturell, gereizt oder gnädiger im Urteil, je mehr Hefte man korrigiert hat.

© Claire Bretécher, 1989

Schüler empfinden daher solche Bewertungen von Leistungen oft als subjektiv und ungerecht.

Da mit Klassenarbeiten in der Schule auf lange Sicht auch immer eine Auslesefunktion verbunden ist, sann man auf Abhilfe. Die Korrektur von Aufsätzen zu standardisieren und damit deren Benotung ve- gleichbarer zu machen, entpuppte sich jedoch als schwierig. Je akribischer ein Korrekturverfa- desto unhandlicher ist es und damit ungee- ete- für die tägliche Unterrichtspraxis. Wie läs- diesem Dilemma entkommen?

Neue psychometrische Testverfa- re-

In den 1960er Jahren wurde die- behavioristischen Lerntheorien und der stru- ra- n Linguistik beruhende audiolinguale -tode- mmer populärer und man schuf der- chtest- dieser Methode ging man davon aus, da- einen be- grenzten Vorrat an gram- ischen -ukturen und Wörtern so eini- kann, - ohne Nachdenken sponta- ind und auf andere Situationen übertrage- rde- en. Voraussetzung dafür ist, dass man j- Struktur genau isolieren und gege- kann. Wie das Lehren kleinste- -iten *(discrete points)* das formale Gerüst der Spra- -hern sollte, so wollte man auch die Fertigkeiten H-ren, Sprechen, Lesen und Schreiben streng getrennt einüben. Tests des Hör- oder Leseverstehens zum Beispiel sollten so konstruiert sein, dass man bei der Lösung nicht auch noch schreiben musste. Der audiolinguale Unterricht w- Gru- die Kunst des kleinschrittigen un- wohlge- -ten Lehrens, das dem Lernen die -ach- Wege bereite- sollte.

-der- -hauptung, Sprachkompetenz resultie-re aus de- einahe -hanischen Addition isolierter Elemente, folgt- Umkehrschluss, dass Lernende, die über die einzelnen Elemente verfügen auch Spra-kompe-enz besitzen müssten. In den 1960er Jahren setzten sich folglich psychometrische Formen der Leistungsmessung durch, die bereits vor dem 2. Weltkrieg entwickelt und erprobt worden waren. Psychometrisch heißen diese Verfahren, weil sie in der Psychologie entwickelt wurden. Mit der Absicht Konstrukte wie Intelligenz, Motivation, Angst usw. zu messen, wurde von beobachtbaren Daten („Igitt, eine Spinne!“) auf Konstrukte geschlossen, die ihnen zu Grunde liegen (Angst, Ekel). Gleichermaßen wollte man von den beobachtbaren Daten (Zahl der richtig gelösten Aufgaben) auf die ihnen zu Grunde liegenden kommunikativen Fähigkeiten schließen: Kann ein Lernender einen Test lösen, so kann man wohl auch annehmen, dass er in der Zielsprache kommunizieren kann.

Man bemühte sich vor allem um die Objektivierung der Leistungsmessung, indem man die verwendeten Tests möglichst zuverlässig (engl. *reliable*) gestalten wollte. Zuverlässig heißt, dass Lernende für eine vergleichbare Leistung auch eine vergleichbare Punktzahl erhalten. Möglichst wenig in einem Test sollte der subjektiven Bewertung des Prüfers überlassen bleiben (siehe Cartoon oben). Des Weiteren

sollten die zu erreichenden Lernziele genau definiert sein und dann objektiv und verlässlich gemessen werden.

Geschlossene Testformate sollten störende Einflüsse bei der Lösung möglichst ausschließen. Die Beurteilung, ob die Ziele erreicht worden waren, sollte nach Möglichkeit kein subjektives Urteil mehr erfordern, sondern neutral und mechanisch erfolgen können. Die *Multiple-Choice*-Aufgabe hielt ihren Einzug in die Testdidaktik:

> The singer ended the concert … one of his most popular songs.
> **A** by **B** with **C** in **D** as

Da die Fremdsprachendidaktik in den 1940er bis 1960er Jahren Sprache zu einem in kleinste Bausteine zerlegbaren Baukasten erklärt hatte, stieß die Isolation von messbaren Aufgaben kaum auf theoretische Bedenken. Die psychometrischen Messverfahren waren schließlich auf dem gleichen Boden einer behavioristischen Psychologie entstanden wie die audiolinguale Methode, was sie für Fremdsprachendidaktiker ohne weiteres akzeptabel machte. Testdidaktiker wie Robert Lado (in seinem epochalen Werk *Language Testing,* 1960) setzten Tests als Standard durch.

Die USA, aber auch Großbritannien, zogen eine sehr große Zahl von Studierenden an, was die Universitäten wiederum vor praktische Probleme stellte, da sichergestellt werden musste, dass diese Studierenden hinreichend gut Englisch sprachen. Aufnahmeprüfungen wurden unumgänglich. Sie mussten vor allem schnell auswertbar sein und zu zuverlässigen Ergebnissen führen. Die Entwicklung einer Testdidaktik großen Stils ist folglich eng mit einem Wirtschaftszweig verknüpft – den Sprachprüfungen großer Organisationen.

In jedem Jahr legen Millionen von Menschen Sprachprüfungen für Englisch ab. Die allermeisten davon wählen eins der großen, international anerkannten Examen der Universität Cambridge (heute *Cambridge English Language Assessment,* dessen bekanntester Test das *Certificate of Profiency in English* ist) oder den amerikanischen *Test of English as a Foreign Language (TOEFL)*, der von Ausländern für eine Immatrikulation an US-amerikanischen Universitäten vorzulegen ist. Man stelle sich vor, welch immense logistische Leistung hier zu vollbringen ist! In *Measured Words* (1995) beschreibt Bernard Spolsky, wie amerikanischen Universitäten Anfang der 60er Jahre den *TOEFL* zu einem maschinell auswertbaren, objektiven, standardisierten und kostengünstigen Prüfungsinstrument entwickelten, um den Korrekturaufwand bei offenen Aufgabenformaten traditionellen Prüfungen zu reduzieren. Mittlerweile ist die maschinelle Auswertung Standard in allen großen Prüfungszentralen, und computerbasierte Sprachtests gibt es inzwischen fast.

Vom psychometrischen zum kommunikativen Test

Schon hatte es bereits unter den Testdidaktikern der 1960er Jahre Auseinandersetzungen um den Sinn dieser psychometrischen Tests gegeben. In seinem Buch *Fundamental considerations in testing for English language proficiency of foreign students* mahnte John B. Carroll (1961) beispielsweise an, dass Tests dieser Art bei komplexeren Fertigkeiten, wie Sprechen und Schreiben nicht sonderlich valide (also inhaltlich aussagekräftig) seien. Selbst wenn ein Kandidat die in kleinschrittige Aufgaben aufgebrochenen Teiltests richtig gelöst hätte, könnte dies nicht unbedingt Auskunft über den wirklichen Stand seiner Sprachkompetenz geben. Dies wurde zum zentralen Kritikpunkt an den Testverfahren der 1960er Jahre und führte nach und nach zu einer terminologischen Änderung: Statt von Tests sprach man jetzt lieber von Assessment und versuchte so, einen neuen Ansatz hervorzuheben.

Im Zuge der Hinwendung zum kommunikativen Fremdsprachenunterricht entwickelte die Testdidaktik in den 1970er Jahren auch kommunikative Sprachprüfungen. Man erkannte, dass kommunikatives Handeln ein komplexes Zusammenspiel verschiedener Teilfertigkeiten und sprachlicher Elemente erfordert, die verständnisvoll und strategisch

planvoll, oder auch kreativ und spielerisch eingesetzt werden. Zwar bemühte man sich auch in kommunikativen Curricula um detaillierte Lernzielkataloge (z. B. im wichtigsten Dokument dieser Zeit, dem *Threshold Level English* des Europarats, entwickelt von Jan van Ek 1975), begriff aber schnell die potenzielle Unendlichkeit der angestrebten Listen sowie ihre relative Beliebigkeit und benutzte sie daher eher als lockere Referenzmittel denn als strenge Lehrpläne.

In kommunikativen Tests ging es wie im Unterricht darum, was man mit Wörtern tun kann, wie man sprachlich handelt, Absichten versprachlicht, Handlungsziele erreicht, sich mit anderen Menschen auseinandersetzt, Texte produziert, austauscht und rezipiert. Formales linguistisches Wissen wurde zwar als notwendige Grundlage für kommunikative Kompetenzen und kommunikatives Handeln akzeptiert, erhielt jedoch im didaktischen Denken eine untergeordnete, quasi dienende Funktion. Als Problem der älteren psychometrischen Tests erachtete man die mangelnde *Validität* (Gültigkeit) der Aufgaben. Kommunikative Praktiker kritisierten, dass die objektivierte Teststruktur ihrer Vorgänger nicht die Lernzielhierarchien, die einem modernen Fremdsprachenunterricht zugrunde liegen. Setzt man sich kommunikativ kompetentes Handeln zum Ziel, muss man überprüfen, ob die Lernenden auch wirklich kommunikativ handeln können. Dies lässt sich aber nur durch Testaufgaben (oder -formate) sicherstellen, die so offen angelegt sind wie menschliche Kommunikation. Um die Paradigmenverschiebung zu markieren, wurde *testing* durch *assessment* und somit zu einem Konzept, dem es um mehr und anderes gehen sollte.

> You recently took part in a class discussion about choosing an interesting job. Your teacher has now asked you to write a composition, answering the following question and giving reasons for your choice.
> Would you rather be a politician, a teacher or a musician?

(ALTE: Materials for the Guidance of Test Item Writers, 1999, S. 139)

Natürlich hatte man dadurch schlafende Hunde geweckt, denn das Problem der (vermeintlichen) Objektivität, das durch die verlässlichen (reliableren) psychometrischen Tests der 1960er Jahre gebannt zu sein schien, stellte sich bei kommunikativen Tests erneut ein.

Der Weg vom behavioristischen zum kommunikativen Paradigma bedeutete gleichzeitig eine Verlagerung des Interesses vom Aspekt der Reliabilität auf den der Validität von Tests. Die beiden Begriffe spiegeln demnach unterschiedliche Konzepte von Sprachtests bzw. Assessment wider.

3 Qualitätsmerkmale von Tests

Die großen Testinstitutionen wie Cambridge English Language Assessment oder der Educational Testing Service mit *TOEFL/TOEIC* haben es geschafft, die inhaltliche Gültigkeit, also die Validität ihrer Tests im Sinne moderner, kommunikativer Ansätze des Fremdsprachenunterrichts, zu aktualisieren. Außerdem tun sie alles, um durch immer raffiniertere statistische Auswertungsverfahren alles Zufällige bei der Bewertung in den Griff zu bekommen und dadurch die Verlässlichkeit von Tests zu verbessern (vgl. Kapitel 13). Von zentraler Bedeutung sind zudem die neueren Ansätze zur Verbesserung der Validität durch klare Orientierung an Dokumenten wie dem *Gemeinsamen europäischen Referenzrahmen für Sprachen: lernen, lehren, beurteilen* (Europarat 2001, vgl. Kapitel 2).

2 Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen

Der *Gemeinsame europäische Referenzrahmen: lernen, lehren, beurteilen* (Europarat 2001, im Folgenden *GeR* abgekürzt) gilt als international wichtigstes Dokument für die Beschreibung von sprachlichen Kompetenzen. Es wäre jedoch zu kurz gegriffen, den *GeR* darauf zu reduzieren, dass er sprachliche Kompetenzen beschreibt und dokumentiert. Primär ist er ein sprachenpolitisches Dokument, das Mehrsprachigkeit befürwortet und für die Förderung von weniger häufig gesprochenen Sprachen eintritt. Für den Fremdsprachenunterri[illegible] be[illegible]nder[illegible] interessant ist auch die Empfehlung für [illegible]ha[illegible]ngsorientierten Ansatz.

Am bekannte[illegible] worden sind allerdings die Skalen [illegible] *Deskriptoren* [illegible]licher Kompetenzen[1]. Die nachhalt[illegible] Wirkung auf das sprachdidaktische Handlungsfeld ent[illegible] die im *GeR* verankerte Skala mit d[illegible]*mzniveaus* A1 bis C2. Die weltweit bekann[illegible]sten [illegible]prac[illegible]tests für Englisch, die auf dieses System [illegible]nt“ [illegible], sind z. B. die folgenden Tests:

A Elementare Sprachverwendung		B Selbstän[illegible] Sprachve[illegible]endung		C [illegible] Kompetente [illegible]prachverwendung	
A1	A2	B1	B2	C1	C2
	Key English Test (KET)	Preliminary English Test (PET)	[illegible] Certificate in [illegible]nglish (FCE)	[illegible]ertificate in [illegible]dvanced English (CAE)	Cambridge Proficiency in English (CPE)

Die Buchstaben A, B und C stehen f[illegible] die [illegible]kann[illegible] Stufen *Beginners* (Grundstufe), *Intern[illegible]* (Mittelstufe) und *Advanced* (Oberstufe). Sie s[illegible] au[illegible] schen Gründen noch einmal unter[illegible] (A1, A2 usw.). Wenn man möchte, kann man im [illegible]ste[illegible] noch weiter differenzieren (B1.1, B1.2 us[illegible]).

Doch wie verhalten sich [illegible]veaus [illegible]nander, und wie leicht lassen sich die sechs [illegible]ehr Stufen erklimmen? Die meisten [illegible]rnation[illegible]en Testanbieter sind sich dari[illegible]ig, dass [illegible]von A1 zu A2 in etwa halb so [illegible]der von A2 bis B1.

> [Die Erfahrung zeig[illegible]ass viele Lernende für den Weg v[illegible]pelt so viel Zeit benötigen wie [illegible]für den [illegible]eg zu [A2]. Sie werden folglich wohl auch mehr als doppelt so lange brauchen, um von [B1] aus [B2] zu erreichen – selbst, wenn die Niveaustufen auf der Skala den gleichen Abstand voneinander zu haben scheinen. Das ist darauf zurückzuführen, dass sich das Spektrum der Sprachaktivitäten, der Fertigkeiten und der sprachlichen Mittel nach oben hin notwendigerweise verbreitert. Diese Tatsache spiegelt sich darin, dass eine Skala von Niveaustufen oft als Diagramm „im Eistütenformat“ dargestellt wird, als ein dreidimensionaler Konus, der nach oben hin breiter wird. Was Aussagen zur durchschnittlichen Lernzeit für das Erreichen eines bestimmten Niveaus mit bestimmten Zielen betrifft, ist größte Vorsicht geboten.
>
> *(Europarat 2001: 29, Kap. 2.2)*[2]

[1] Seit einigen Jahren ist eine Neubearbeitung auf dem Weg, die 2017 abgeschlossen wurde. Einige dieser Änderungen werden auch in Kapiteln unseres Buches aufgegriffen.

[2] Des Textflusses wegen zitieren wir nur aus der gedruckten deutschen Fassung, die 2001 im Langenscheidt-Verlag erschienen ist. Die „elektronische“ Version findet man auf der Website des Goethe-Instituts unter http://www.goethe.de/Z/50/commeuro/deindex.htm (letzter Zugriff: 01.12.2017).

Weil inzwischen immer mehr Menschen als politische Flüchtlinge oder aus wirtschaftlichen Gründen nach Europa kommen und die jeweilige Landessprache nie in der Schule lernen konnten, gibt es in der neuen Fassung des *GeR* (2018) eine Stufe *pre-A1* (dt. vor A1), also eine Art *survival level*. Aber auch für die Grundschule oder für dritte Fremdsprachen könnte dieses Niveau interessant sein. Diese Erweiterung um pre-A1 wird wahrscheinlich zu einer Umstrukturierung des ganzen Systems der Referenzniveaus führen, die neu „geeicht" werden müssten.

Wie die Beschreibu[illegible] so[illegible]er [illegible]fen im Detail aussieht, zeigt die urs[illegible]ch[illegible] Globalskala aus dem Jahr 200[illegible]der alle anderen Skalen abgeleitet sind.

Kompetente Sprachverwendung	**C2**	*Kann* praktisch alles, was er/sie liest [illegible]hört, mühelos *verstehen*. […] *Kann sich* spontan, sehr flüssig und genau[illegible]drücken und auch bei komplexeren Sachverhalten feinere Bedeutungsnu[illegible]en deutlich machen.[3]
	C1	*Kann* ein breites Spektrum anspruchsvoller längerer Texte *verstehen* und auch implizite Bedeutungen erfassen. *Kann* die Sprache im gesellschaftlichen und beruflichen Leben […] und flexibel gebrauchen. […] *Kann sich* klar, strukturiert und ausführlich zu komplexen Sachverhalten *äußern*. […]
Selbstständige Sprachverwendung	**B2**	*Kann* die Hauptinhalte komplexer Texte zu konkreten und abstrakten Themen *verstehen*; versteht im eigenen Spezialgebiet auch Fachdiskussionen. *Kann sich* so spontan und fließend *verständigen*, dass ein normales Gespräch mit Muttersprachlern ohne größere Anstrengung auf beiden Seiten gut möglich ist. *Kann* zu einem breiten Themenspektrum klar und detailliert [illegible]en, […]
	B1	*Kann* die Hauptpunkte *verstehen*, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. *Kann sich* einfach und zusammenhängend über vertraute Themen und persönliche Interessengebiete *äußern*. […]
Elementare Sprach-verwendung	**A2**	*Kann* Sätze und gebräuchliche Ausdrücke *verstehen*, die mit Bereichen von ganz unmittelbarer Bedeutung zusammenhängen (z. B. Informationen zu Person und zur Familie, Einkaufen, Arbeit, nähere Umgebung). *Kann sich* in einfachen, routinemäßigen Situationen *verständigen*, in denen es um einen einfachen und direkten Austausch von Informationen über vertraute und alltägliche Dinge geht.
	A1	*Kann* vertraute, alltägliche Ausdrücke und ganz einfache Sätze *verstehen* und verwenden, die auf die Befriedigung konkreter Bedürfnisse zielen. […] *Kann sich* auf einfache Art *verständigen*, wenn die Gesprächspartnerinnen oder Gesprächspartner langsam und deutlich sprechen […].

(Europarat 2001: 35, hier[illegible] gekü[illegible]v d. Verf.)

In der G[illegible]hl rezeptive als auch produktive [illegible]tenzen aufgeführt. Um den Aufbau und die Str[illegible]einzelner Deskriptoren zu verdeutlichen, wurden „Kann" und das darauf folgende Verb kursiv hervorgehoben.

Diese Skala wird sich wegen des neuen Niveaus pre-A1 ein wenig verändern. Durch „vor A1" verschieben sich auch zwangsläufig einige Elemente im gesamten Gefüge. Auf diesem neuen Niveau findet sich nun beispielsweise eine Variante des ursprüng-

[3] Vor allem auf Niveau C2 enthält die Ergänzung von 2017 zum *GeR* einige Änderungen: wurden im *GeR* 2001 noch „Muttersprachler" erwähnt, die man verstehen kann, heißt es jetzt „Gesprächspartner", die flüssig und schnell sprechen.

lichen Deskriptors auf A1, der den Sprecher so einstuft, dass er Fragen über sich selbst stellen und beantworten kann, die unmittelbare Bedürfnisse und Routinen zum Inhalt haben und sich dabei formelhafter Redewendungen oder gar Gesten bedienen kann, um die Information zu unterstützen. Solche Bedeutungsunterschiede sind sehr schwer an konkreten Versprachlichungen festzumachen, weil die Unterschiede zwischen „vor A1" und A1 oft minimal sind. Zumal sie für den Alltag des Assessments auch wenig bedeutsam sind, da sich Klassenarbeiten in der 5. Jahrgangsstufe eher am Lehrbuch orientieren werden als an Kompetenzen, über die ein Tourist oder Zuwanderer verfügen sollte.

Weitere Verästelungen [illegible]lb dieser Globalskala finden sich in un[illegible]geordneten Skalen zu den speziellen Fertigkeiten [illegible]ören, Sprechen usw. Diese ausführ[illegible]ersionen sind präziser als die übergeordnete [illegible]kala. Exemplarisch soll dies an der [illegible]la des *GeR* [illegible]ündlichen Interaktion gezeigt wer[illegible]d zwar beschränkt auf die Niveaus A1 bis B2.

	Mündliche Interaktion allgemein
B2	Kann die Sprache gebrauchen, um flüssig, korrekt [illegible]d wir[illegible]n über ein breites Spektrum allgemeiner, wissenschaftlicher, beruflicher Them[illegible] oder über Freizeitthemen zu sprechen und dabei Zusammenhänge zwischen Ideen deutlich [illegible]
B1	Kann Informationen austauschen, prüfen und [illegible]tätigen, [illegible] weniger routinemäßigen Situationen umgehen und erklären, warum etwas proble[illegible]tisch ist. Kann Gedanken zu eher abstrakten kulture[illegible]emen ausdrücken, wie z. B. zu Filmen, Büchern, Musik usw.
A2	Kann ohne übermäßige Mühe in einfa[illegible]inegesprächen zurechtkommen; kann Fragen stellen und beantworten und in vorhers[illegible]bare Alltagssituationen Gedanken und Informationen zu vertrauten Themen austauschen.
A1	Kann sich auf einfache Art v[illegible]tätigen, [illegible] die Kommunikation völlig davon abhängig, dass etwas langsamer wiederholt, [illegible]muliert oder korrigiert wird. Kann einfache Fragen stellen [illegible]n, einfache Feststellungen treffen oder auf solche reagieren, sofern es sich um unmittelbar[illegible] Bedürfnisse oder um sehr vertraute Themen handelt.

(Europarat 2001: 114, Auszug A1 – B2)

Das System der Referenzniveaus [illegible] eine [illegible]hebliche Dynamik entfaltet. Alle [illegible]n Test[illegible]eter in Europa (und viele weltweit) habe[illegible] Sprachprü[illegible]fungen auf den *GeR* abg[illegible]stimmt.[4] D[illegible]ses *Common European Frame[illegible]k of Re[illegible]r Languages: Learning, Teach[illegible]ment (CEFR)*, wie es im Original heißt (C[illegible]il [illegible]pe 2001), ist ein Hilfsmittel, mit dem [illegible] vor allem wichtige Prüfungen b[illegible]n und „eichen" kann. Ein großer [illegible] dieses [illegible]odells liegt darin, dass sich bei der Beu[illegible] unterschiedliche Profile in den Kompetenzen der Lernenden erfassen lassen: Einige von ihnen können z. B. flüssig lesen, andere hören und sprechen besser.

Eine ‚dritte Dimension'

Im *GeR* fallen in erster Linie die Skalen mit den Deskriptoren für sprachliche Aktivitäten (Sprechen, Hören, Lesen, Schreiben usw.) ins Auge. Sie bestehen aus einer horizontalen Dimension (den Kompetenzbeschreibungen) und einer vertikalen Dimension (den Stufungen von „vor A1" über A1 bis C2). Die Skalen basieren in erster Linie auf Textsorten, Themen und Inhalten und behandeln den jeweiligen Umgang damit. Hierfür braucht man allerdings vielfältige Mittel. Dazu zählen Wortschatz, Syntax, eine verständliche Aussprache, aber auch interkulturelles Wissen und das Verfügen über sozial angemessene Ausdrucksweisen. Diese Katego-

[4] Eine Übersicht findet sich auf http://www.alte.org/membership (letzter Zugriff: 01.12.2017).

rien bilden eine Art ‚dritte Dimension', die zu der Erfassung von Sprachkompetenz beitragen.

Zu dieser ‚dritten Dimension' gibt es ebenfalls Skalen, mit deren Hilfe man sprachliche Kompetenzen grob beschreiben kann. Zur Illustration werden die Deskriptoren der Niveaus A1 und B1 im Hinblick auf die Verwendung von Grammatik angeführt.

	Grammatische Korrektheit
B1	Kann sich in vertrauten Situationen ausreichend korrekt vers[illegible]; im Allgemeinen gute Beherrschung der grammatischen Strukturen trotz deutlicher E[illegible] der Muttersprache. Zwar kommen Fehler vor, aber es bleibt klar, was ausgedrück[illegible]erden soll.
A1	Zeigt nur eine begrenzte Beherrschung einiger weniger einf[illegible]grammatischer Strukturen und Satzmuster in einem auswendig gelernten Repertoire.

(Europarat 2001: 114, Auszug A1 und B1, stark gekürzt)

Es gibt zwar keine Listen mit Wortschatz oder Syntax, mit deren Hilfe man Texte einstufen könnte, aber ein nachträglich entstandenes Werkzeug: *English Profile*[5] hilft bei der Analyse von Texten und hat sich auch für das Erstellen von Tests als nützlich erwiesen.

Im Zus[illegible]hang mit dieser ‚dritten Dimension' wi[illegible]oft üb[illegible]en, dass beim Verstehen von Tex- [illegible] un[illegible]deren Pro[illegible]tion auch soziolinguisti- s[illegible]rat[illegible]che und pragmatische Kompetenzen benötigt [illegible]den. Am einfachsten sind *soziolinguistische Kompetenzen* zu erläutern. Im Fall von Niveau [illegible]ißt es beispielsweise im *GeR*:

B1	Kann ein breites Spektrum von Spra[illegible]unk[illegible]ner realisieren und auf sie reagieren, indem er/sie die dafür gebräuchlichsten Redemitt[illegible]in n[illegible]rales Register benutzt. Ist sich der wichtigsten Höflichkeitsfo[illegible]sst und handelt entsprechend. Ist sich der wichtigsten Unt[illegible]de zwischen den Sitten und Gebräuchen, den Einstellungen, Werten und Überzeugungen [illegible]treff[illegible]den Gesellschaft und in seiner eigenen bewusst und achtet auf entsprechende Signale.

(Europarat 2001: 122)

Zu den Kompetenzen in die[illegible]itten D[illegible]ension' zählt außerdem Handlungskom[illegible] *(pragmati[illegible]sche Kompetenz)*, die sich am Beisp[illegible]daran zeigt, wie man Gespräc[illegible]organis[illegible]Wort ergreifen, zurückfrag[illegible]. Wichtig sind weiterhin *strategische Kompe[illegible]zen*, [illegible]men auch Rezeptionsstrategi[illegible]für das [illegible]n und Hören von Texten in einer [illegible]t werden.

Strategische Kompetenzen umfassen u. a. die Fähigkeit zum Umschreiben oder das Erbitten von Erklärung. Im *GeR* werden die strategischen Kompetenzen der Niveaus von A2 bis B2 folgendermaßen beschrieben:

[5] Auf der Homepage http://www.englishprofile.org/ (letzter Zugriff: 01.12.2017) werden Englischlehrkräften zahlreiche Hilfen und Materialien angeboten. Ein Clip auf der Internetseite sowie über https://youtu.be/L1-HREEDa70 (letzter Zugriff: 01.12.2017) gibt einen guten Überblick über das Angebot.

B2	Kann eine Vielfalt von *Strategien* einsetzen, um das Verstehen zu sichern; dazu gehört, dass er/sie beim Zuhören auf Kernpunkte achtet sowie das Textverständnis anhand von Hinweisen aus dem Kontext überprüft.
B1	Kann in Texten mit Themen aus dem eigenen Fach- oder Interessengeb[iet] [unbe]kannte Wörter aus dem Kontext erschließen. Kann die Bedeutung einzelner unbekannter Wörter aus dem Kontext [erschl]ießen und Satzbedeutung ableiten, sofern das behandelte Thema vertraut ist.
A2	Kann sich eine Vorstellung von der Gesamtaussage kurzer Text[e] [mit] Äußerungen zu konkreten, alltäglichen Themen machen und die wahrscheinliche Bed[eu]tung unbe[kannter] Wörter aus dem Kontext erschließen.

(Europarat 2001: 78, Auszug A2 – B2)

Solche Formulierungen unterstreichen noch einmal, dass der *GeR* ein *mehrdimensionales Modell von Kommunikation* ist, das Kompetenzen beschreibt und stuft, indem es eine erste und zweite Dimension, aber auch pragma- und soziolinguistische Kompetenzen, die bei der Kommunikation eingesetzt werden.

3 Wozu man die tool box GeR benutzen kann

Nicht nur in Deutschland verbreitete sich der *GeR* wie ein Lauffeuer: Lehrbuchverlage erkannten blitzschnell den Wert für ihr Marketing, Volkshochschulen und auch das allgemeinbildende Schulwesen griffen vor allem die sechs Referenzniveaus und ihre inhaltlichen Füllungen, die Deskriptoren der Sprachkompetenzen, begierig auf. Ein Grund für die Attraktivität des *GeR* beim Lesen war, dass mit der kommunikativen Wende des Fremdsprachenunterrichts seit den 70er Jahren viele kompetenzorientierte Arbeitsformen Einzug in den Unterricht gehalten hatten. Klassenarbeiten und Abschlussprüfungen aber der Entwicklung hinterher hinkten.

Defizite des deutschen Lehrens wurden durch Tests wie etwa dem *TOEFL*[6] aufgedeckt, und Studentinnen und Studenten mussten sich fragen, was das Abitur eigentlich über ihre Sprachkenntnisse aussagte. Auch Abschlüsse für die Mittlere Reife verlangten immer stärker nach einer Objektivierung und Vergleichbarkeit ihrer Ergebnisse. Vor allem die *PISA*-Studien, in denen die Bundesrepublik in der Lesefähigkeit stets nur mittlere Plätze erreichte, schreckten die Bildungspolitik auf. Der *GeR* war daher ein willkommenes Mittel, um auch den Fremdsprachenunterricht in übersichtliche Systeme mit transparenten Abschlussprüfungen zu bringen (vgl. auch Kapitel 14).

Die sechs Referenzniveaus des *GeR* wurden von der Ständigen Konferenz der Kultusminister (KMK) als Basis für die Gestaltung der neuen Bildungsstandards benutzt. Der *GeR* ist also so etwas wie die Mutter aller Kompetenzskalen. Die folgenden Stufungen bzw. Vorstellungen davon, welche Kompetenzniveaus Lernende in der Regel bis zu einem bestimmten Zeitpunkt erreichen sollten, kristallisierten sich in den Köpfen der meisten Lehrerinnen und Lehrer – insbesondere aber bei den Schulverwaltungen – heraus:

Grundschule	A1 (bei 6 Jahren A2)
Hauptschule, 9. Klasse	A2
Mittlerer Schulabschluss	B1
Abitur	B2 bzw. C1

[6] *Test of English as a Foreign Language*, vom Educational Testing Service in Princeton, USA, seit 1964 angeboten.

Kapitel 4

Hör- und Hör-/Sehverstehen

1 Geeignete Hör- und Lesetexte finden

Dass Validität ein zentrales Merkmal von Assessment ist, wurde bereits in Kapitel 2 thematisiert. In diesem Zusammenhang wurden Hilfsmittel wie der *GeR* betrachtet, die vor allen Dingen zur Beschreibung inhaltlicher Validität nützlich sind. In Kapitel 3 stand dann die Konstruktvalidität im Zentrum. Zu deren Erläuterung wurden andere Skalen und Deskriptoren des *GeR* hinzugezogen, in denen ebenfalls Prozesse und Strategien beim Hör- und Leseverstehen beschrieben sind. In diesem Kapitel soll nun beides miteinander verknüpft werden, wobei wird ein weiterer Begriff zur Beschreibung von Validität eine Rolle spielen: der der „Authen

Bei der Suche nach geeigneten Hör- und Lesetexten sind Bildungsstandards oder der [illegible] manchmal nicht sehr hilfreich, da sie im Unterschied zu [illegible] onelleren Curricula Listen mit Textsorten enthalten. Die Deskriptoren des *GeR* sind zudem nicht sehr systematisch formuliert. Sie erwähnen zum einen Texte und Situationen, die man verstehen soll, zum anderen die Art der Sprache, die dabei benutzt wird, oder die Bedingungen und Beschränkungen, unter denen das geschieht, die Relevanz dieser Merkmale variiert allerdings von Stufe zu Stufe.

Die folgenden Ausschnitte aus der Skala ‚Hörverstehen allgemein' liefern einen Eindruck davon. Zunächst soll der Fokus auf den Themen des Hörverstehens liegen, danach geht es um die Anforderungen, die beim Verstehen der Sprache gestellt werden. In beiden Fällen wurden die Deskriptoren durch Kursivsetzung hervorgehoben:

	Hörverstehen Allgemein – Themen	**Hörverstehen Allgemein – Art der Sprache**
C1	Kann genug verstehen, um längeren Redebeiträgen über *nicht vertraute abstrakte und komplexe Themen* folgen zu können, wenn auch gelegentlich Details bestätigt werden müssen, insbesondere bei fremdem Akzent. [...]	Kann genug verstehen, um längeren Redebeiträgen über nicht vertraute abstrakte und komplexe Themen zu folgen [...], *insbesondere bei fremdem Akzent.* Kann ein *breites Spektrum idiomatischer Wendungen und umgangssprachlicher Ausdrucksformen* verstehen und Registerwechsel richtig beurteilen. […]
B1	Kann unkomplizierte *Sachinformationen über gewöhnliche alltags- oder berufsbezogene Themen* verstehen und dabei die Hauptaussagen und Einzelinformationen erkennen, sofern artikuliert und mit vertrautem Akzent gesprochen wird.	Kann unkomplizierte Sachinformationen über gewöhnliche alltags- oder berufsbezogene Themen verstehen und dabei die Hauptaussagen und Einzelinformationen erkennen, sofern *klar artikuliert und mit vertrautem Akzent gesprochen* wird.
A1	Kann [*was genau?* die Verf.] verstehen, wenn sehr langsam und sorgfältig gesprochen wird und wenn lange Pausen Zeit lassen, den Sinn zu erfassen.	Kann verstehen, wenn *sehr langsam und sorgfältig gesprochen wird* und wenn *lange Pausen Zeit lassen*, den Sinn zu erfassen.

(Europarat 2001, S. 71 f.)

Ohne erkennbare Systematik wird mal dieses und mal jenes Merkmal betont. Das macht die Entscheidung nicht einfacher, ob ein gewählter Text für ein Assessment geeignet ist. Ähnliches lässt sich auch bei den Bildungsstandards beobachten. Zur Diskussion der Konstruktvalidität soll daher das Konzept der Authentizität herangezogen werden.

2 Authentizität

Obgleich die meisten Menschen über die Aussprache des Begriffs stolpern, ist er in der Fremdsprachendidaktik unverzichtbar. Wörterbücher paraphrasieren den Begriff *authentic* oft mit *real, genuine* oder auch *known to be true*. Didaktisch greifen diese Umschreibungen jedoch zu kurz. Daher wird in der Fachdidaktik regelmäßig gefordert, dass die Sprache, die in Hörtexten Verwendung findet, der von realen Sprechsituationen oder Texten entsprechen soll. Damit wird allerdings ausgeblendet, dass auch Originaltexte, sobald sie für Lehrwerke oder Tests ausgewählt (und eventuell auch gekürzt oder bearbeitet) sind, nicht mehr völlig authentisch sind.

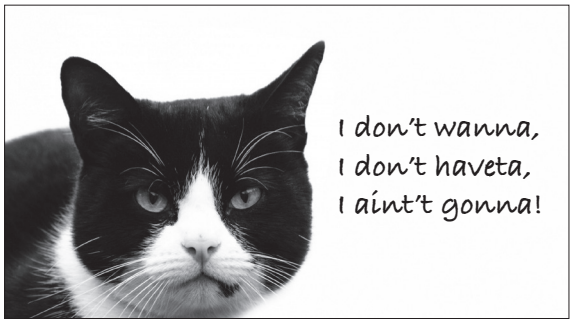
Aus der Forderung nach Authentizität ergeben sich eine Reihe von Qualitätskriterien: Erstrebenswert wäre beispielsweise, dass Hörtexte nicht nur von Lehrbuchautoren oder Testkonstrukteuren für didaktische Zwecke erstellt werden. Sie sollten vielmehr aus natürlichen Medien stammen, die für Muttersprachler produziert und von Muttersprachlern gelesen oder gesprochen werden. Finden lassen sich authentische Hör-/Sehtexte problemlos bei Funk und Fernsehen, in Mediatheken, bei Streaming-Diensten usw.

Texte, die für native speakers erstellt sind, haben zudem nicht die *Eigenschaft* authentisch zu sein, sie eröffnen auch authentische Möglichkeiten des Umgangs mit ihnen: Einen Film sieht man sich ihn. Später diskutiert man eventuell mit Freunden darüber. Im Unterricht hingegen beantwortet man Lehrerfragen dazu oder macht Inhaltsangaben, bevor man mit anderen Lernenden über den Sinn diskutiert (natürlich unter Anleitung der Lehrkraft, die das steuert). Letzteres wäre also kein authentischer Umgang mit einem Text mehr, sondern ein „didaktischer".

Nun lässt sich argumentieren, dass alle Unterrichts-Sprechsituationen künstlich sind und *„willing suspension of disbelief"* voraussetzen (der Begriff stammt schon von 1817 von Samuel Taylor Coleridge geprägt). Hinzu kommt, dass authentische Hörtexte im Unterricht oft auf Ablehnung stoßen, weil sie den Schülern zufolge zu schnell gesprochen wird, Hintergrundgeräusche und Dialekte das Verständnis erschweren usw. Didaktisch vereinfachte Tonaufnahmen der Schulbuchverlage dominieren folglich den Unterrichtsalltag und führen dazu, dass viele Jugendliche erst in der Oberstufe authentisches Englisch zu hören bekommen. Dank des eigenen Medienkonsums haben sie bis dahin aber schon so manche Varietät des Englischen zu hören bekommen und bestenfalls auch verstehen gelernt.

Gesprochene Sprache zu verstehen ist ein wichtiges Ziel im kompetenzorientierten Unterricht. Authentische Sprache klingt aber ganz anders als die didaktisierte Sprache in bereinigten Tonaufnahmen, bei denen es durchaus passieren kann, dass der Regisseur einen Sprecher zur Wiederholung einer Passage auffordert: *„John, can you read that again? There was a stomach rumble."*

Neben britischem und amerikanischem Englisch gibt es Dialekte und andere Varietäten, Auslassungen und Verschleifungen. Denken Sie beispielsweise an *fish'n'chips, bread'n'butter, lots of* (gesprochen /'lɑtsəv/) oder den „Peng!cake" (= *pancake*).



© *Pixabay/PublicDomainPictures*

Bei mündlichen Äußerungen gibt es Brüche in der Syntax und Füllsel zur Überbrückung von Pausen *(erm, well)*, in Dialogen gibt es Überlappungen und vieles mehr. Einmal ganz abgesehen davon, dass Menschen nicht im sorgfältig abgeschirmten Schallraum eines Studios sprechen, sondern ihre Stimme gegen Hintergrundgeräusche anheben, die Intonation auch emotional steuern etc. Wenn wir Lernende auf de

3 Hörtexte finden und bearbeit

Für die Arbeit mit Hörtexten ist z al, e Vorstellung davon zu haben, welch sorten für die jeweilige Zielgruppe relevant nd in Zukunft sein werden. Eine Stu g des Schwierigkeitsgrads könnte dann mith d oder der Bildungsstandards erfolgen. E chw gkeit bleibt aber dennoch beste enaufn en für Lehrwerke und Tests sind in de l nicht au thentisch. Wenn sie es nahmsw se sind, entsprechen sie mei cht den ines Kompetenzmodells wi

Semi-authentisc He ial, das die Lernenden noch nicht ke und das den Vorzug hat, auf besti ht zu sein, findet man bei den gro tanbiete n:

Dschungel gesprochener Sprache vorbereiten wollen, ist der Einsatz authentischer Hör- und Sehverstehenstexte im Unterricht und beim Assessment unumgänglich. Nur so l die Kompetenzen der Lernenden entwick un ber ufen.

Leider ist Authentizi m noch kein Qualitätskriterium n vollem Umfang akzeptiert wird, weder be chve lagen noch bei Testanbiete n. Man muss genau hinhören und prüfen, wie s ch Höraufnahmen zumindest um „Semi-Authentizi hen, und das auch schon auf nied Niveaustufen. Da gibt es durchaus gradue Un sch de. Im Grunde kann man aber schon z en s , wenn man neben den Höraufna en zu erken für Assessment überhaupt t xte det, die in „natürlichem" Sprechtempo gesp hen sind, mit altersgemäßen Stimmen und zumi st einige der oben genannten Merkmale authentischer Sp he widerspiegeln.

(1) Für einen Test auf dem Niveau A2, dem *Key English Test*[1] von Cambridge English Language Assessment. Auf dieser Website kann man sich auch zu allen anderen Niveaustufen durchklicken und sich eine Vielfalt von Aufgaben herunterladen, vor allem auch Hördateien im MP3-Format.
(2) TELC = The European Language Certificates[2].
(3) *DIALANG* (eine vom Europarat angebotene online Ressource)[3].
(4) *TOEFL*/*TOEIC*[4].

Diese leicht zu erschließenden Hörmaterialien, die nicht nur auf *CEFR*/*GeR*-Niveaustufen angeboten werden, sondern dazu noch mit Aufgaben versehen

[1] http://www.cambridge.org e/cambridgeenglish/catalog/cambridge-english-exams-ielts/key-schools/resources (letzter Zugriff: 01.12.2017).

[2] https://www.telc.net/pruefungsteilnehmende/sprachpruefungen.html (letzter Zugriff: 01.12.2017).

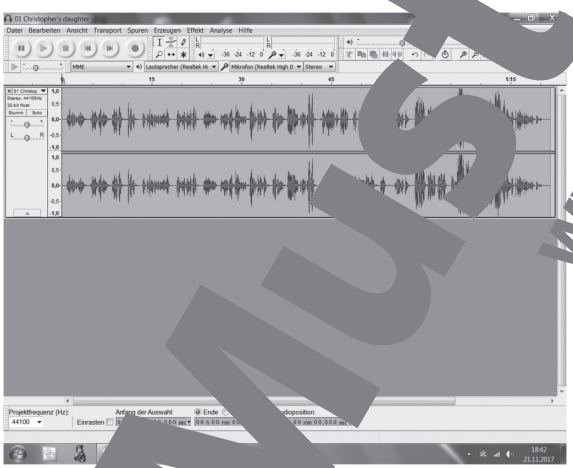
[3] http://www.lancaster.ac.uk/researchenterprise/dialang/about (letzter Zugriff: 01.12.2017): Die Nutzung ist etwas kompliziert, unter der angegebenen Adresse gibt es aber eine ausführliche Anleitung.

[4] http://www.ets.org/toefl/ibt/prepare/quick_prep/ (letzter Zugriff: 01.12.2017).

sind bilden eine ideale Quelle für eigenes Assessment im Bereich Hörverstehen.

Wenn man authentische Materialien für mittlere und obere Lernstufen sucht, sollte man die vielen Angebote von britischen oder US-amerikanischen Radio- und Fernsehsendern nutzen, bei denen man allerdings mit zusätzlichen Anforderungen konfrontiert wird. Es gilt nicht nur Hörtexte zu finden, sondern diese danach auch noch zu bearbeiten. Für beide Herausforderungen folgen nun einige Quellen und Hinweise für die Aufbereitung.

Es gibt eine außerordentlich nützliche Website der BBC mit einer großen Auswahl an Hörtexten,[5] die sich für den Einsatz im Englischunterricht anbieten und besondere für das *GeR*-Niveau B1 und höher eignen. Auch bei YouTube gibt es Podcasts.[6] Für ihre Nutzung sollte man sich allerdings mit einer Bearbeitungs-Software vertraut machen, da diese Texte von der Länge her nicht unbedingt für den Unterricht geeignet sind. Mit *Audacity*, einem kostenlosem Audio-Editor, der mit den bekannten Audio-Formaten zurechtkommt (MP3, WAV u.a.), geht das ganz einfach.



Screenshot einer in Audacity importierten Hördatei.

Sobald das Programm auf dem Rechner installiert wurde, ist es möglich, Audiodateien zu importieren und zu bearbeiten. So lassen sich beispielsweise einzelne Teile markieren und ausschneiden, die Klangeigenschaften können verändert, Höhen und Tiefen beeinflusst oder aber eigene Aufnahmen professionell durch Ausblenden bearbeitet werden. Die fertig bearbeitete Audiodatei kann schließlich auf einer CD, einem USB-Stick oder in der *Cloud* gespeichert und dann im Unterricht abgespielt werden.

Nützlich ist die Bearbeitung insbesondere bei vielen YouTube-Videos oder Mitschnitten von Podcasts. Bei YouTube-Videos benutzt man zunächst einen *Converter* wie *Freemake Free YouTube Downloader* oder vergleichbare Freeware-Programme. Daneben braucht man in der Regel einen *MP4 to MP3 Converter*, der auch kostenlos im Internet erhältlich ist. Dieser ist notwendig, um die im MP4-Format enthaltenen Videos in Hördateien zu konvertieren. Findet man also z.B. einen interessanten Vortrag oder bei dem die visuelle Komponente nicht so wichtig ist, kann man das MP4-Video in eine MP3-Hördatei umwandeln und dann mit *Audacity* weiter bearbeiten (siehe Screenshot). Wenn das übersichtlich gespeichert und allen im Kollegium zugänglich gemacht wird, beteiligen sich wahrscheinlich auch immer mehr Personen am Suchen und Bearbeiten solcher Hörtexte. Auf diese Weise gelangt man schnell zu einem Pool an Höraufnahmen, den man sowohl für unterrichtliche Zwecke als auch bei der Erstellung von Klassenarbeiten nutzen kann.

[5] www.bbc.co.uk/podcasts (letzter Zugriff: 01.12.2017).

[6] https://www.podcastchart.com/categories/top-200-podcasts?page=1#position-4 (letzter Zugriff: 01.12.2017) oder auch www.itunescharts.net/charts/podcasts/ (letzter Zugriff: 01.12.2017).

Kapitel 9

Schreiben beurteilen

1 Geschlossene Aufgabenformate

Bei der Beurteilung von schriftlichen Leistungen ist zunächst relevant, ob es sich um geschlossene oder offene Aufgabenformate handelt (vgl. Kapitel 3). Die Wahl der Aufgabenformate hängt wiederum von unterschiedlichen Faktoren ab. So spielt das Kompetenzniveau der Lernenden eine Rolle: Wieviel Sprache können sie zusammenhängend produzieren? Auch die Vorgaben von Lehrplänen müssen bedacht werden: Welche Kompetenzen sind besonders wichtig (z. B. schriftliche Produktion oder Interaktion)? Schließlich müssen die Anforderungen an die Lernenden berücksichtigt werden: Welche Textsorten erwarten den Lernenden im späteren Leben begegnen? Welche Kompetenzen müssen die Lernenden in beruflichen Situationen bewältigen? Letztlich ist eine Frage der Zuverlässigkeit von Bewertungen, für welche Aufgabenformate man sich entscheidet. Geschlossene Aufgabenformate *(True/False, Multiple Choice etc.)* erlauben größere Zuverlässigkeit, also *Reliabilität* (vgl. Kapitel 13).

2 Bewerten von schriftlichen Leistungen mit offenen Aufgabenformaten

Bei der Bewertung offener Aufgabenformate steht eher die Validität im Vordergrund. Zentral ist die Zielgerichtetheit einer Aufgabe und die Leitfrage: Misst die Aufgabe die Kompetenz, die geprüft werden soll? Dadurch, dass Bewertende eine Entscheidung über die Qualität mittels Sprache gelösten Aufgabe treffen müssen, leidet die Reliabilität möglicherweise. Schließlich haben Entscheidungen immer ein subjektives Moment, sodass bei der Bewertung von Aufgabenformate keine vollkommene Verlässlichkeit gegeben ist. Man kann aber sehr wohl dafür sorgen, dass die Beurteilungsprozesse transparent sind.

Die Formen der Bewertung bei offenen Aufgaben sind denen ähnlich, die in Kapitel 7 vorgestellt wurden. Ebenso wie beim Sprechen können Checklisten oder Beurteilungsraster verwendet werden, um geschriebene Lernertexte zu beurteilen. Dabei eignen sich insbesondere für Lernende mit geringen Vorkenntnissen Checklisten: Sie können mit wenig Aufwand auf eine bestimmte Aufgabenform abgestimmt werden und eignen sich auch zum Einsatz bei *self* und *peer assessment*. Die Checkliste (vgl. S. 102) bezieht sich auf eine per E-Mail ausgesprochene Einladung an einen Freund. Die Checkliste wird zuerst als Instrument für *self assessment* verwendet und dann für die unterrichtsbasierte Leistungsbeurteilung seitens der Lehrkraft.

Häufig kommen bei der Bewertung von produktiven und interaktiven schriftlichen Leistungen in der Fremdsprache auch Beurteilungsskalen zum Einsatz. In standardisierten Tests sind sie schon seit langem Usus, wenn es um die Bewertung schriftlicher Leistungen geht. Bei landesweiten Abschlusstests setzen sie sich ebenfalls durch. In der Funktion eines Erwartungshorizontes bzw. als Bewertungsraster sind sie darüber hinaus in manchen Bundesländern bei Klassenarbeiten Pflicht. Im Grunde unterscheiden sich die Beurteilungsskalen für Sprechen und

	You		Your teacher	
Important points	Yes ☺	No ☹	Yes ☺	No ☹
I have put in a greeting (Dear …).				
I have asked how my friend is (How are you?).				
I have invited my friend to come to my hometown.				
I have written about interesting places or things to do in my hometown.				
I have asked him/her what s/he likes to do best.				
I have ended the e-mail with a greeting (Love, All the best etc.).				
I have written in a friendly tone.				
	You:		Your teacher:	
Comments/improvements:				

Writing an e-mail to a friend-Checkliste

Schreiben nur in einem Punkt, nämlich in ihrem Umfang. Beurteilungsskalen für das Schreiben können etwas detaillierter sein, weil der geschriebene Text im Gegensatz zum Sprechen vorliegt und von Beurteiler mehrmals gelesen werden kann.

Für die Verwendung von Skalen [illegible] erfordert die Wahl zwischen *holistischen* Skalen, die eine Leistung ganzheitlich bewerten und keine sprachlichen Kriterien explizit ausweisen (aber dennoch welche beinhalten) und *analytischen* Skalen, die die Bewertungskriterien einzeln auswerten und mehrere Teilergebnisse in diesem Verfahren zu einem Endergebnis zusammenfügen (vgl. auch Kapitel 7). Es lässt sich eine fertige Skala konsultieren, deren Kriterien der Bewertung entsprechend übernommen oder angepasst werden müssen. Alternativ kann aber auch eine eigene Skala mit den Bewertungskriterien, die als wichtig erachtet werden und mit den jeweiligen Zielen übereinstimmen, erstellt werden.

Das folgende Beispiel (vgl. S. 103) stammt von der *Association of Language Testers* in Europe (ALTE 2005) und ist eine holistische Skala, die sich völlig aufgabenunabhängig verwenden lässt.

Der Vorteil dieser Bewertungsanleitung liegt darin, dass man sie für alle Aufgabentypen – auch über das Schreiben hinaus – verwenden kann. Darüber hinaus lässt sich neben sprachlichen Aspekten auch die kommunikative Wirkung der Äußerung bewerten, sodass die Lernenden angemessen beurteilt werden können. Deren Sprache weist zwar die eine oder andere Unebenheiten auf, sie trauen sich aber beispielsweise, komplexere Sprache zu verwenden, um eine Aufgabe auf kreative Weise zu lösen. Damit wird die Bewertungsanleitung interessant für aufgabenbasierte Beurteilung.

5	Aufgabe voll erfüllt. • Alle Inhaltspunkte werden in angemessenem Umfang behandelt. • Große Breite der Strukturen und des Wortschatzes gemäß der Aufgab • Minimale Fehler, vielleicht weil zu ehrgeizig; gut entwickelte Beherrsch g d pr che. • Effektiver Aufbau der Gedanken mit einer Vielfalt verbindender Eleme • Register und Format durchgehend der Schreibabsicht und dem Les r an Erzielt voll die gewünschte Wirkung auf den Leser.
4	Gute Erfüllung der Aufgabe. [...]
3	Ausreichende Erfüllung der Aufgabe. • Alle hauptsächlichen Inhaltspunkte werden behandelt; e re Auslassungen. • Angemessene Breite der Strukturen und des Wo chatzes, die die Anforderungen der Aufgabe erfüllen. • Eine Reihe von Fehlern ist möglich, sie behinder n t die Ko munikation. • Angemessener Aufbau der Gedanken mit ei hen Verbi ungen. • Ausreichender, wenn auch nicht immer erf eicher Ver uch, ein der Schreibabsicht und dem Leser angemessenes Register und Format v wenden. Erzielt im Großen und Ganzen die gew Wirku g auf de Leser.
2	Versuch, die Aufgabe zu erfüllen, der n t a messen gelingt. [...]
1	Schwacher Versuch, die Aufgab • Erhebliche inhaltliche Au sun und/oder häu irrelevante Inhalte, möglicherweise wegen Nicht-Verstehen der Au Anweis • Geringe Breite des Wortsc zes der Str turen. • Häufige Fehler, die d mmuni on behi rn; kaum entwickelte Beherrschung der Sprache. • Kein Aufbau der Gedan ine Verbind gselemente. • Kein oder nur ge ges Bew ein von Register und Format. Die Äuße eine s gative Wirkung auf den Leser.

Holistische Skala zum S n (7: 12)

Für geschriebene Texte speziell auf mittlerem Niveau, beispielsweise für die 8. bis 10. Klasse, hat Charlotte Stolle in ihrer Arbeit für das hessische Staatsexamen (2009) ein interessantes Beurteilungsraster entworfen, das in vereinfachter Form aufgegriffen werden soll.

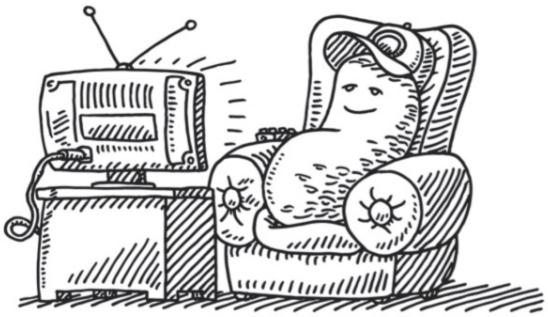
Kriterium	Unterteilung	Gewichtung	
1. Gesamteindruck			
	Kohärenz: Form, Inhalt, Eigenständigkeit etc.	3/25	
2. Inhalt			
	• Quantität (evtl. laut Vorgabe)	3	7/25
	• Qualität	3	
	• sachliche Richtigkeit	1	
3. Adressatenbezug			
	• formale Aspekte (korrekte Anrede, Höflichkeit oder Vertrautheit etc.) • inhaltliche Aspekte (Vorwissen etc.)	4/25	
4. Sprache			
	• Korrektheit	3	7/25
	• Komplexität	3	
	• Angemessenheit Textsortenbezug	1	
5. Gliederung			
	• Einteilung von Absätzen • Stimmigkeit	4/25	

Beurteilungsraster für die Mittelstufe (nach Stolle 2009).

Stolle integriert in ihrem Raster kommunikative und linguistische Aspekte, beispielsweise sinnvolle Satzanfänge, eine idiomatische Ausdrucksweise, die Verwendung von Konnektoren wie „und" bzw. „aber" und eine differenzierte themenangemessene Wortwahl. Das Raster hat den Vorteil, dass es gleich eine Gewichtung der Punkten vornimmt, die man natürlich auch nach eigenen Schwerpunkten ändern oder erweitern kann.

Das folgende Beispiel stammt aus einer Realschulabschlussprüfung:

Ausgehend von einem visuellen Impuls und drei Leitfragen *(What are the reasons for being a couch potato? Are you a couch potato or do you know a person who is a couch potato? Do you enjoy being a couch potato in your free time?)* sollen die Schüler etwa 150 Wörter zum Thema produzieren, sei es über sich selbst, eine/n Freund/in oder allgemein. Eine Textsorte ist nicht vorgegeben.



The Couch Potato Family (unkorrigiert)

My family is a idol for all the others couch potatoes in the world. I don't know why it's so. I expect that my father works the whole day and in the evening he is very tired. My mother doesn't want that the family go outside. She is a real couch potat because she doesn't work and she does nothing the whole day, apart from the housework am the only one in the family who do a sp At the weekends my parents sometimes Frankfurt or in Rheingau. Then I et fri or my boyfriend. We go shopp in the cinema.

I do much sport with my frien to mple is video-clip dancing or athlet Back o my family. Perhaps I can und d the less by my father, because he works ve But I don understand the lazine my m her. She sits the whole day ome a o her that we can go shopp sn't want. That's my lazy family. (173 wor

Geht ma spiels das Raster von Stolle (200 h, sieht man, dass nicht alle Kriterien für diese etwas unspezifische Textsorte gleichermaßen hilfreich sind. Die Kategorie „Gesamteindruck" ist unproblematisch, weil sie recht allgemein ist und tex ezifische Schwerpunktsetzungen zuläss Der halt t den Komponenten „Quantität" un bli Richtigkeit" ist ebenfalls hanc Lediglich die Kategorie „Qualität" würde ver eine Diskussion unter den Bewert en im Leh gium ergeben, etwa um sich dar verständigen, ob die Leitfragen im Schülertext al ksichtigt werden müssen. Das ist i Beispiel nicht der Fall, obwohl der Text in tlich rot lem kohärent ist.

Prob sch st in diesem Fall das Kriterium „A essate g". Formale Aspekte wie die angesse ede oder e passendes Register fallen in Gewicht. Ähnlich verhält es sich mit der inhal en Adress tenangemessenheit. Da der Leser für diesen Te kein Vorwissen benötigt und die Textso nicht spezifiziert war, lässt sich über dieses K erium wenig sagen.

Aufgrund dessen fällt die Punkteverteilung schwer. Beim Kriterium Sprache stehen die Korrektheit und die Komplexität (im *GeR* in etwa unter „Spektrum der sprachlichen Mittel" gefasst) sowie die Angemessenheit im Vordergrund. Folglich sind bereits viele Aspekte einer sprachlichen Leistung abgedeckt. Für die Beurteilung etwas umfangreicherer Texte in geschriebener Form ist zudem die Kategorie „Gliederung" hilfreich, da sie den Bewertenden helfen, die stimmige Struktur und den Aufbau des Texts sowie dessen Kohäsion und Kohärenz einzuschätzen.

Ohne die entsprechenden curricularen Vorgaben zu kennen, die dieser Leistung zugrunde liegen, ist es sicherlich schwierig, individuelle Leistungen zu bewerten. Ein Vorschlag zur Einschätzung der Schülerarbeit wäre:

Kriterium	Unterteilung	Gewichtung		Bewertung
1. Gesamteindruck				
	Kohärenz: Form, Inhalt, Eigenständigkeit etc.	3		3
Kommentar	Recht eigenständige Leistung mit geschickt gewähltem Anfang und Ende.		5	
2. Inhalt				
	• Quantität (evtl. laut Vorgabe) Erfüllt die Zielvorgabe von 1500 Wörtern und ist umfangreich genug, um eine Gesamtaussage zu entwickeln.			3
	• Qualität Sie beantwortet die Leitfragen zwar nicht vollständig, aber bringt eine pointierte, leicht ironische personalisierte Darstellung zur Meinung der Verfasserin inklusive eigenständig formulierter Vorwürfe an die Adressaten.		7/25	2
	• sachliche Richtigkeit Bei der gewählten Textsorte (persönliche Meinungsäußerung) kann man eigentlich nichts falsch machen.	1		1
3. Adressatenbezug				
	• formale Aspekte (korrekte Anrede, Höflichkeit oder Vertrautheit etc.) • inhaltliche Aspekte (Vorwissen etc.)	3		3
Kommentar	Diese Kategorie ist schwierig, möglicherweise ist der Adressatenbezug in Ordnung, wenn die Verfasserin eine neutrale Adressierung wählt. Ansonsten erfordert der Text kein Vorwissen.		3/25	
4. Sprache				
	• Korrektheit Unebenheiten in der Wortstellung (Einfluss des Deutschen auf den Satzbau) und einige kleinere grammatische Fehler, die das Verständnis nicht beeinträchtigen:	3		2
	• Komplexität Die Verfasserin verwendet einfache Strukturen und wenig komplexes Vokabular, dass sich auch wiederholt:	3	7/25	1
	• Angemessenheit Textsortenbezug Die gewählte Textsorte der personalisierten Meinungsäußerung ist angemessen und schlägt sich in einem [illegible]isch gefärbten Stil nieder:	1		1
5. Gliederung				
	• Gliederung von Absätzen • Stimmigkeit	4/4	4/25	4
Kommentar	Der Text ist trotz Redundanzen kohärent und hat einen gelungen gewählten Anfang und Ende.			

Bewertungsorschlag: 20/25 Punkten.

Die Alternative zu „fertigen“ Beurteilungsrastern ist eine selbst erstellte Skala. Hier besteht die Möglichkeit, die Kriterien zur Beurteilung auszuwählen, die jeweils wichtig sind: Auf diesem Weg lässt sich u. U. eine bessere Passung zwischen den Lernzielen des Unterrichts und der Bewertung erzielen. Außerdem kann die Skala auf ein spezielles Aufgabenformat hin abgestimmt werden. Das ist gleichzeitig Vorteil und Nachteil in einem: Die Bewertung lässt sich zwar genau auf die Aufgabe anpassen, dies führt aber dazu, dass erstens für jedes Aufgabenformat eine neue Skala zu erstellen ist (Zeitaufwand) und zweitens auf der Grundlage der Bewertungsergebnisse keine Rü[illegible]e auf die allgemeine Kompetenzentwicklung [illegible] der [illegible]em [illegible]prache möglich sind. Passend zur Aufgab[illegible]ch [illegible]ato“ wurde eine „selbstgestrick[illegible]che Bew[illegible]eilungsskala mit den Bewertungskrite[illegible]che Inhalt, Kohärenz und Aufgabe[illegible]füllung erstel[illegible]

		1	2	3	4
Sprache	Spektrum der sprachlichen Mittel				
	Adressatenangemessenheit				
	Korrektheit				
Inhalt	Korrektheit/sachliche Richtigkeit				
	Kreativität/Ideenreichtum/Origin[illegible]t				
Kohärenz	Aufbau/Gliederung/logische St[illegible]tur				
Aufgabenerfüllung	Kommunikativer Erfolg/kom[illegible]kative Aufgabe erfolgreich erfüllt				

Beurteilungsskala zu „Couch Potato“

Die Skala ist zwar auf die Aufgabe abges[illegible] würde aber auf verschiedene Aufga[illegible]nf[illegible]mate[illegible] treffen. Eine explizite Gewichtung d[illegible]ebnisse mit Punkten gibt es nicht, es ist aber [illegible]mög[illegible] Zahlen als Punktzahlen oder all[illegible]einer als Niveaustufen zu betrachten. Bei [illegible]zte[illegible]iante kann das Ergebnis direkt in eine [illegible]„üb[illegible]etzt“ werden, wenn man noch ein[illegible]te hinz[illegible]gt.

Ob holistisch oder analytisch, [illegible]n auf ein[illegible] Kompetenz oder eine [illegible]olle Au[illegible]abe: Beurteilungsraster bzw. [illegible]teilung[illegible]ffen, die Subjektivität der B[illegible]en offenen Aufgabenformaten zu reduzie[illegible]und [illegible]Transparenz (wenn auch nicht [illegible]nbedingt [illegible]Reliabilität) zu erhöhen. Darüber [illegible] gutes Instrument zur Rückmeldu[illegible]die Leis[illegible]ng der Lerner dar. Insbesondere bei an[illegible]en Skalen lassen sich Kompetenzprofile erstellen, [illegible]ie es erleichtern, den Lernenden Auskunft über deren aktuelle Fähigkeiten zu geben als auch auf Möglichkeiten zur Verbesserung, also wie die nächstmögliche Niveaustufe erreicht werden kann, hinzuweisen.

[illegible]m Weiterlesen

[illegible]ür die[illegible]en [illegible]bschnitt gelten die gleichen Empfehlungen wie [illegible]ür Kapitel 8:

Weigle, Sara Cushing. (2002). *Assessing Writing*. [illegible]ambridge: Cambridge University Press.

Weigle, Sara Cushing. (2012). Assessing Writing. In: Coombe, Christina/Davidson, Peter/O’Sullivan, Barry/Stoynoff, Stephen (Hrsg.). *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press. 218–223.

6 Unverzichtbar: das Beurteilertraining

Dass selbst die Benutzung von Beurteilungsbögen und Bewertungsrastern nicht immer eine hinreichende Objektivität garantiert, wurde bereits thematisiert. Deshalb empfiehlt es sich, auf Fachkonferenzen und mit Hilfe von gemeinsamen Korrekturen nach einem vorher diskutierten Beurteilungsraster ein größtmögliches Einvernehmen über Schülertexte zu erzielen. Abweichende Einschätzungen müssen dabei diskutiert werden, aber nicht mit dem Ziel, eine Kollegin oder einen Kollegen unbedingt zu überzeugen, sondern besser zu verstehen, wo die Gründe für die Differenz liegen: Vielleicht wurde das Korrekturraster von einer Kollegen anders ausgelegt oder missverstanden? Der Nutzen von Beurteilertraining für die *inter-rater reliability* (Übereinstimmung zwischen verschiedenen Beurteilern) ist in zahlreichen Untersuchungen nachgewiesen worden. So lästig es sein mag: ein Beurteilertraining ist unverzichtbar, wenn man in hohem Maße verantwortungsvoll mit Schülerleistungen umgehen will.

7 Geschlossene Aufgabenformate und das Kriterium Reliabilität

Wenn man eine größtmögliche Zuverlässigkeit (Reliabilität) bei Bewertungen erreichen möchte, bieten sich die im ersten Kapitel erwähnten geschlossenen Aufgabenformate (in der Regel *multiple choice*) an. Eine nicht zu unterschätzende Schwierigkeit zeigt sich dabei oft schon bei der Konstruktion solcher Aufgaben. Eingangs wurde bereits auf das Problem der Ratewahrscheinlichkeit bei sollen Mehrfachwahl-Aufgaben hingewiesen. Diese Wahrscheinlichkeit sollte man nicht auf die leichte Schulter nehmen.

Die Konstruktion von Aufgaben zu Strukturen und Wortschatz ist es oft schwer, für eine einfache binäre Entscheidung *(„X or Y?" – „Present progressive or past ?")* eine richtige Lösung und vier Distraktoren zu schreiben, die alle gleich plausibel sind und von denen nicht ein oder zwei von den Lernenden als offenkundig falsch durchschaut werden. Das würde nämlich das Ergebnis eines Assessments verfälschen.

In professionellen Testsystemen gibt es daher – stets nach der Erprobung oder dem eigentlichen

Aufgabenanalyse (nach Elisabeth Ingram 1968)[1]

Anzahl der Schüler: 25 (: 3 =) 8 8

Item	oberes Drittel		mittleres Drittel		unteres Drittel		richtige Lösungen	% (1)	Differenz (2)	Σ 1 – 3 (3)
1	///// ///	8			///// ///	8	25	100	0	0
2	///// //		////	4		0	12	48	8	1
3	///// /			5	///	3	14	56	3	.38
4		8	///// ///	8	///// /	6	22	88	2	.25
5			///	5	//	2	14	56	5	.63
6	///		/	1	///// //	7	11	44	- 4	- .50
7	///		//	2	/	1	6	24	2	.25

(1) Alle richtigen Lösungen geteilt durch Anzahl der Schüler x 100 (in Item 3 also 14 : 25 x 100 = 56 %)
(2) Oberes minus unteres Drittel (in Item 3 also 1: 6 - 3 = 3)
(3) Differenz oberes/unteres Drittel geteilt durch die Anzahl der Schüler geteilt durch 3 (in Item 3 also 3 : 8 = .38)

[1] Eine Vorlage der Aufgabenanalyse nach Ingram lässt sich auf der Homepage des Helbling Verlags zum Download finden.

Test – eine Aufgabenanalyse. Ein einfaches Verfahren schlägt Ingram vor. Nach einer Vorkorrektur aller Arbeiten (bei der alle Lösungen erfasst werden) ermittelt man für alle Schüler, hier 25, welche Arbeit wie gut ausgefallen ist. Daraufhin sortiert man die Hefte oder Lösungsbögen in absteigender Reihenfolge und unterteilt sie in ein oberes, mittleres bzw. unteres Drittel. Dann legt man für jedes Item eine Strichliste an und hält fest, wie oft Schüler des oberen Leistungsdrittels die Aufgabe richtig gelöst haben und wie oft das im mittleren bzw. unteren Drittel der Fall war. In der Tabelle (S. 134) ist solch eine Strichliste für 7 Aufgaben erfasst und ermöglicht durch die statistische Auswertung einige interessante Beobachtungen.

Interpretation der Werte

Richtige Lösungen / in %

Dieser Wert verrät den Schwierigkeitsgrad eine… Items. Zu einfache (d. h. 90 – 100% richtige), und auf… alle Fälle zu schwierige (= unter 20% richtige) … sollte man aus der Bewertung herausnehme… wei… beide niedrige Trennschärfewerte haben und … nicht viel über Leistungsunterschied… aus…en.

∑ 1 – 3 (= Trennschärfe): Ideal… hier ein… Wert von + 1 (wie in Item 2), der zeigt, …

Item genau zwischen den guten und den schwächeren Schülern unterscheidet. Inakzeptabel sind der Wert 0 (wie in Item 1, das von guten und von schwächeren Schülern … gelöst wurde) oder negative Werte (wie in … das … von den schwächeren Schülern besser … wurde. Professionelle Testkon…ure schließen Items mit Werten unter .35 m…

Lehrkräfte s… sich ein- oder zweimal im Schuljahr die Mühe ma… ihre selbst erstellten Tests einer so… fung zu unterziehen, um ein Gefühl dafür z… entw…keln, welche Aufgaben zu leicht oder zu schw… sind, u… welche bei der Beurteilung von Sc…lerleist…, sowohl für die Lehrkraft als auch … die Schüler, von a… messener Schwierigkeit s… eich… gilt für Tests oder Lernerfolgskontrollen aus L…werken, die auch nicht immer von Testexperten erstellt …rden. Bei der professionellen …struktion von Testaufgaben sollte es selbstverst…ch sein, dass diese erprobt werden und dass m…ssglückte …ms aus dem Test getilgt werden. Professionell erstellte Aufgaben-/Itemanalysen geben … zusätzl…ch noch über weitere Dinge Auskunft.

Iteman

Item and Test Analysis Program – …AN…n) Version 3.50

Seq No.	Scale -Item	Prop. Corr…	D… Index	Point …iser	A…native	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
8	2-1	.38	…58	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					C	.12	.11	.12	-.01	
					D	.49	.74	.23	-.44	
					Other	.01	.00	.00	-.11	
9	2-2	…71	.42	.42	A	.07	.11	.01	-.16	
					B	.11	.18	.04	-.22	
					C	.10	.16	.00	-.22	
					D	.71	.53	.95	.42	*
					Other	.01	.00	.00	-.13	

(ALTE Materials for the Guidance of Test Item Writers. Module 2, 2005: 61 (gekürzt))

Hier sind mithilfe eines Computerprogramms die Aufgaben/Items 8 und 9 ausgewertet worden. *Scale-Item* bedeutet die Nummer des Items in einem bestimmten Subtest (Prüfungsteil, wie z. B. Lese- oder Hörverstehen), und der Prozentsatz der korrekten Antworten *(Proportion Correct)* war bei 8 = 38% und bei 9 = 71%. Die Trennschärfe *(Discrimination Index)* lag für 8 bei .52, für 9 bei .42, war also in beiden Fällen zufriedenstellend.

Point Biserial ist eine andere Form der Berechnung der Trennschärfe, und auch hier sind bei beiden Items die Werte zufriedenstellend. Bei dieser gründlicheren Aufgabenanalyse wurde zusätzlich noch der Wert der vier Wahlmöglichkeiten A bis D berechnet (* = die [illegible]Antwort). Bei der Auswertung kann man also [illegible]sehen, welche der Antwortmöglichkeiten [illegible]end oft gewählt wurde, was au[illegible]ute Qualität der Distraktoren schließen lässt. [illegible]fgabe 8 wurde zum Beispiel die Ant[illegible]t A überhaupt [illegible]t gewählt, sollte also bei einer er[illegible] Nutzung der Aufgabe neu formuliert werden.

8 Fazit

Während man früher vor allem in Prüfungen der renommierten Testanbieter allergrößten Wert auf Objektivität und Zuverlässigkeit (Reliabilität) von Tests legte, ist man heute eher bemüht, kompetenzorientiertes sprachliches Assessment sinnvoll (valide) und trotzdem auch verlässlich zu gestalten. Assessments in der Klasse sollte man sich e[illegible]fall[illegible] um hohe Qualitätsstandards bemühen, wenn aber auch einmal ein Auge zudrücke[illegible] Thomas von Aquin hat geschrieben, [illegible]Gerechtigkeit ohne Gnade Grausamkeit sei. Lehrkräften daher nie vergessen, dass di[illegible]ligen Schüler, die in diesem Versetzungszeugn[illegible]no[illegible]einer Note 4 davongekommen sind, im [illegible]sten[illegible]r die Fünferkandidaten sein werd[illegible]schden[illegible]e Fünfer des laufenden Jahres die Klass[illegible]sen haben. Und da Lehrkräfte – im [illegible]ensatz [illegible]den großen kommerziellen T[illegible]anbieter[illegible]ngeschränkte Möglichkeiten [illegible]ssicherung haben, empfiehlt es sich, ruhig[illegible]ch [illegible]ig Barmherzigkeit im Sinne de[illegible] Thomas [illegible] Aquin walten zu lassen.

[illegible]m V[illegible]erlesen

[illegible]200[illegible]2007). *Materials for the Guidance of Test Item[illegible]riters./Hi[illegible]reichungen für Testautoren.* [illegible]ttp://www.alte.org/[illegible]sources/filter (letzter Zugriff: [illegible]2017).

[illegible]Empfehlenswert, wenn man über Grundkenntnisse der Materie verfügt und für Kundige eine aktuelle und zuverlässige Informationsquelle.

Henning, Grant (2013). *A Guide to Language Testing: Development, Evaluation, Research*. Boston, Mass.: Heinle & Heinle. Erste Auflage 1987.

Einführung in statistische Verfahren bei Sprachtests; die Neuauflage dokumentiert die Popularität und Aktualität dieses „Klassikers".

McNamara, Tim (2000). *Language Testing*. Oxford University Press.

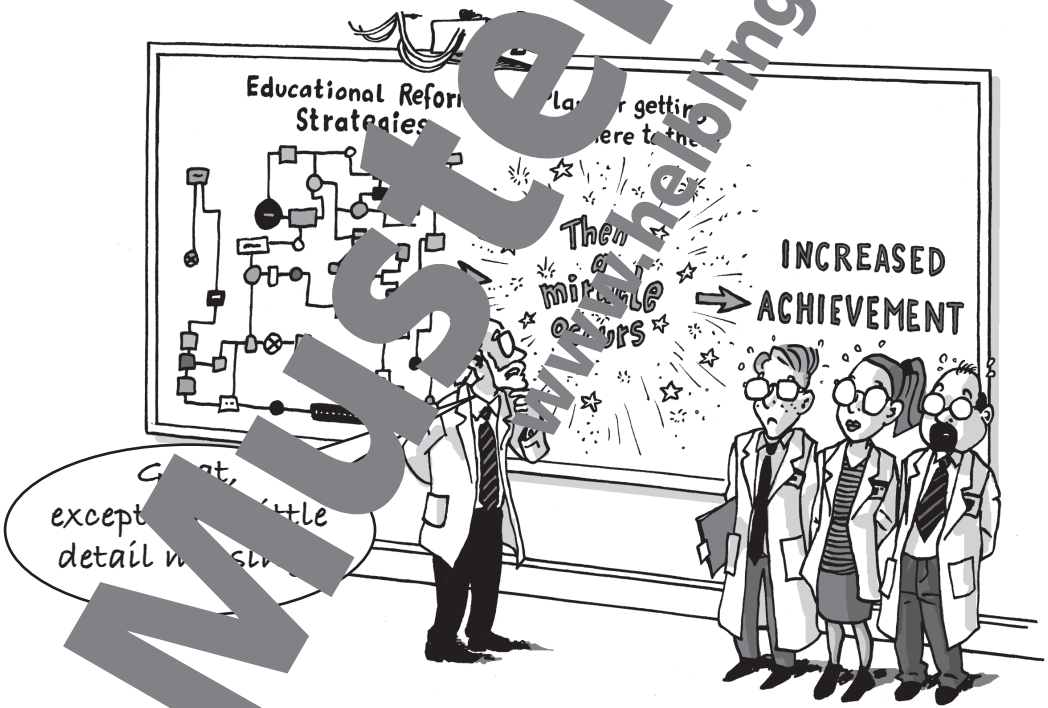
Erläutert neuere Verfahren der statistischen Prüfung bei Sprachtests wie *Item Response Theory*, das aktuell in der statistischen Auswertung von empirisch erhobenen Daten (nicht nur bei Tests!) favorisierte Verfahren. IRT ist keine einzelne Theorie, sondern eine Familie von formalen, mathematischen, probabilistischen Messmodellen, welche postulieren, dass dem beobachtbaren Testverhalten (manifeste Variable) eine bestimmte Fähigkeit / Eigenschaft bzw. Disposition (latente Variable) zugrunde liegt. Mehr dazu bei McNamara.

Kapitel 14

Vom Wiegen wird die Sau nicht fett

„Vom Wiegen wird die Sau nicht fett“ war ein beliebter Spruch von Lehrkräften, denen zentrale Lernstandserhebungen (von der Grundschule bis zum Abitur) und andere Eingriffe in ihren Unterricht den Nerv raubten. Das sei „Testeritis“! Statt den Erfolg von Unterricht messen zu wollen, solle man sich lieber um eine Verbesserung der Unterrichtsbedingungen kümmern.

Zentrale Prüfungen haben aber durchaus wichtige Funktionen jenseits des eigenen klasseninternen Assessments, auf die im Folgenden eingegangen werden soll. In diesem Zusammenhang wird immer wieder der Begriff des „Bildungsmonitorings“ fallen. Mit diesem Begriff ist eine Leistungsüberprüfung des deutschen Bildungssystems gemeint, auf das sich die Kultusministerkonferenz im Jahr 2006 geeinigt hat, um systemische Defizite auszumachen und zu beheben. Als Instrument wurden u. a. Vergleichsarbeiten (*VERA*) zur Überprüfung landesweiten und [illegible] Leistungsfähigkeit beschlossen.



© *Ulf Marckwort*

Bei *„Educational Reform Strategies“*, so der Text des Cartoons, möchte man sich nicht auf Wunder verlassen, sondern die Steuerung auf solide Daten stützen und dazu zählen natürlich möglichst objektive Messungen, was Schüler gelernt haben und können. Dies ist eine in angelsächsischen Ländern selbstverständliche Überlegung, in Deutschland war das bis *PISA* ungew[illegible]neu.

1 PISA und die Folgen: Bildungsstandards

PISA wird eigentlich in Großbuchstaben geschrieben, da es für *Programme for International Student Assessment* steht und nicht für die Stadt mit dem schiefen Turm. Es ist ein Versuch der OECD, die Kenntnisse und Kompetenzen von Fünfzehnjährigen in Mathematik und Naturwissenschaften sowie ihre Lesefähigkeit in der jeweiligen Muttersprache zu ermitteln. Die OECD handelte im Auftrag von Regierungen, in Deutschland der Kultusministerkonferenz, um „Bildungsmonitoring“ zu betreiben, also nicht nur eine „Ist-Beschreibung“ der Ressou[illegible]cen zu liefern, die einer Gesellschaft in Zukunft zur Verfügung stehen, sondern auch zu deren Ve[illegible]serung beizutragen. Warum also der Aufsc[illegible]ei[1] in den Medien und unter den Lehrkräften?

Getroffene Hunde bellen, hei[illegible]e Vo[illegible]weisheit. *PISA* entzauberte die deu[illegible] Schule, die sich zuvor unter den besten Bildungss[illegible] der Welt gesehen hatte, denn di[illegible]undesrepublik schnitt in dieser Studie anfangs m[illegible]mi[illegible]ig ab. In der Folge wollte man zum ei[illegible]genau wissen, wo das deutsche Bildu[illegible]sen in[illegible]ational steht und zum anderen, wie s[illegible]zelne Bun[illegible]desländer im Vergleich [illegible]reinan[illegible] behaupten konnten. Dahint[illegible]eckten [illegible]e partei- und bildungspolitisc[illegible]en, wie etwa die Frage nach der Struktu[illegible]s [illegible]stems (Stichwort: Gesamtschule) und d[illegible]förderung bestimmter sozialer Gr[illegible]der, Mädchen, später auch Kinde[illegible]Migranten). Vor allem „katholische Mädchen au[illegible]terfamilien auf dem Lande“ waren seit den 1965er Jahren, als Georg Picht in Deutschland die „Bildungskatastrophe“ ausrief, der Prototy[illegible]örderungswu[illegible]chülergruppen. Dies sah man 30 J[illegible]äter als behoben an. Doch dann kam die Kritik de[illegible]D.

Man [illegible]ich zu fragen, was Finnland und andere [illegible]kan[illegible]avis[illegible]e Länder in ihrem Schulwesen bes[illegible]ach[illegible] als die Bundesrepublik. Insbeso[illegible]äre im[illegible]ich der Lesekompetenz lagen sie [illegible]en [illegible]chen Lände[illegible]an der Spitze der Vergl[illegible] Al[illegible]olge dieses „*PISA*-Schocks“ schuf die Kultusmi[illegible]erkonfe[illegible]z „Bildungsstandards“ sowie zentrale Absch[illegible]sprüfungen, mit deren Hilfe [illegible] Analysen a[illegible]f den Prüfstand gestellt werden soll[illegible].

Das rief die Kritiker noch stärker auf den Plan – *PISA* selbst: Bildung könne man nicht standardisiere[illegible]tete ein zentraler Vorwurf. Interessanterweise geschah dies oft unter Berufung auf den [illegible]mboldtschen Bildungsbegriff aus dem 19. Jahrhundert. Bildungsforscher wie Eckhard Klieme versuchten die Debatte zu versachlichen, indem sie begriffliche Präzisierungen vorschlugen:

> Bildungsziele fallen nicht vom Himmel und sie haben nicht den Status unbefragbarer Gewissheiten, sondern verdanken sich historischen Kontexten und nationalen Traditionen. Wer „Allgemeinbildung“ sagt, der beansprucht eine – erkennbar deutsche – Tradition des „Bildungs“-denkens und der Interpretation von „Kultur“, [...] wer von „Basisfähigkeiten“ ausgeht, nimmt die Debatte über Standards und notwendige Erwartungen an Schule zur Kenntnis. Bildungsziele sind also, sichtbar an solchen Traditi-

[1] Die Diskussion um *PISA* ist in Wikipedia sehr gut dokumentiert: https://de.wikipedia.org/wiki/PISA-Studien (letzter Zugriff: 01.12.2017)

> onen, in ihrer konkreten Gestalt immer Ergebnis gesellschaftlicher Entscheidungen und sozialer Machtlagen [...].
> *(Klieme 2003: 58)*

In den Chor der Kritiker stimmten auf der „Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts" auch die meisten Fremdsprachendidaktiker der deutschen Hochschulen ein (Bausch et al. 2003). Mit „Standardisieren" waren dabei nicht nur mögliche Stufungen gemeint, sondern die Profile von Abschlussprüfungen. Es waren nämlich just die Tests von Cambridge Assessment, die seit dem Erscheinen des *Gemeinsamen europäischen Referenzrahmens für Sprachen* (Europarat 2001) auf diesen „geeicht" waren. Als Musteraufgaben wurden sie zunächst in den Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) veröffentlicht: im Jahr 2003 für den Mittleren Bildungs

abschluss und im Jahr darauf auch für den Hauptschulabschluss.

Im Grunde war klar, was auf Lehrkräfte (und Lernende) zukam. Begei ste dies allerdings nicht gerade aus. Di Cam dg Prüfungen unterschieden sich erheb n gängigen Klassen- und Abs beiten. ent, dass diese herkömmlichen Ar nen höheren Bildungswert hatten die Aufgab in standardisierten Tests benutz den. Das übersah man großzügig. Viel schwerer fie ewicht, dass Cambridge-, TOEFL- re Tests Alltagstexte und Alltagssituati n h anz en, um Lese- und Hörverstehen zu prüf Trotz der „kommunikativen W de" de te noch immer die Grammatik, hren Entwicklu g und Überprüfung von K en erst ga z langsam im Fremdsprachenunte nt Einzu halt.

2 DESI: Bildungsmonitoring für en nglischunterricht

Parallel zu den Bildungsstandards g di KMK nem Konsortium von Fachdidaktikern ildungsforschern den Auftrag, ergänzend z Pi Bestandsaufnahme in den Fächer Deutsch und Englisch zu versuchen: *DESI = De ch- Schülerleistungen-International*.[2] Dies wa erst nlauf zum Bildungsmonitoring für nglisch erricht, mit dem sich Lehrkräfte konfron hen, nach dem Deutsch und Math ik scho Bei *PISA* auf dem Prüfstand w Die ses Bildungsmonitorings wa end. Am Ende der neunten Jahrgangsstufe ich) zwei Drittel der Schüler im Mündliche s Ge Niveau A2, das die KMK für chluss erwartete. Ein Drittel erre B1, das f den Mittleren Schulabschluss erwartet Allerdings erreichten sogar neun Prozent B2, wie es eigentlich erst in der Oberstufe erwartet werden kann. Ähnliches gilt für andere Teilfertigkeiten.

> DESI zeigt damit, dass wir im Englischunterricht – vor allem in den Gymnasien – eine sehr starke Leistungsspitze von 10 bis 15 Prozent der Schülerinnen und Schüler haben, deren Kompetenzen weit über das Anforderungsniveau der Lehrpläne hinausragen. Andererseits zeigt DESI vor allem in Hauptschulen, Integrierten Gesamtschulen und Schulen mit mehreren Bildungsgängen deutliche Defizite. Im Bildungsgang Hauptschule erreicht etwa nur ein Drittel der Schülerinnen und Schüler das Regelziel der Bildungsstandards.
> *(Klieme 2006: 2)*

[2] Der Name war etwas hoch gehängt, denn „International" war die Studie nur, weil auch einige Schulen in Österreich und die deutschsprachigen Schulen in Südtirol miteinbezogen wurden.

Des Weiteren zeigten die Ergebnisse, dass die fremdsprachlichen Leistungen von Mädchen denen von Jungen überlegen sind. Oft waren auch Kinder mit Migrationshintergrund denjenigen Kindern überlegen, die Deutsch als Muttersprache sprechen:

> Das Aufwachsen in einer mehrsprachigen Familie ist unter sonst gleichen Lernbedingungen (sozialer Hintergrund, kognitive Grundfähigkeiten, Geschlecht, Bildungsgang) im Englischen mit einem Leistungsvorsprung verbunden, der den Gewinn mindestens eines halben Schuljahres ausmacht. Auch Schülerinnen und Schüler, die aus Migrationsfamilien mit ausschließlich nicht-deutschem Sprachhintergrund stammen, zeigen im Englischunterricht vergleichsweise gute Leistungen.
> *(DIPF 2006: 5)*

Hinzu kam ein weiterer Befund, der nicht unbedingt anderen Studien entsprach: Sprachliche Kompetenzen (auch in Englisch) stehen laut *DESI* eher im Zusammenhang mit [illegible]rellen Ressourcen der Familie und damit [illegible]ie [illegible]sitiv [illegible]ie Familie Bildung allgemein gegenü[illegible]br[illegible]d nicht unbedingt mit dere[illegible]mmen.

Mit Aktionen [illegible]SA un[illegible] *DESI* werden Lehrkräfte i[illegible]er häufiger [illegible]tiert, sodass sich ein Gewöhnung[illegible] einstellte. Daher halten sich die abwehrenden Rea[illegible] darauf mittlerweile auch in Gren[illegible] gilt ebenfalls für die vom IQB vorgelegte[illegible]udi[illegible]ildu[illegible]gstrend[3], die seit 2015 jährlich vorgele[illegible]d u[illegible]ie weiteren Bildungsberichte.

3 „Is my B1 your B1?“

Mit welchen Problemen sich Bildungsforsch[illegible] aus[illegible] einandersetzen müssen, kann man anhand [illegible] Studien beobachten, in denen es [illegible] Erm[illegible]lung des Niveaus von Fremdsprach[illegible]tnissen in Europa ging: dem *European Survey [illegible]n L[illegible] Competences* (2012) und der Fol[illegible]studie *Study on Comparability of Language Testin[illegible] F[illegible]* 2015). Obwohl Deutschland an der 2012e[illegible]die [illegible] nicht beteiligt war und auch die 2[illegible] Studi[illegible]cht das deutsche Schulsystem betraf, lo[illegible]ein kurze[illegible] Exkurs, um Einblicke in [illegible]methoc[illegible]en Schwierigkeiten zu gewi[illegible]

Beiden Unt[illegible]lagen die Kompetenzniveaus des *GeR* [illegible]und[illegible]ts, die sich daran orientierte[illegible] sollten [illegible] Sprachkompetenzen von Schülerin[illegible]überprüfen.[4] In beiden Fällen ging [illegible]r ander[illegible]n um eine Antwort auf J. Charles Alders[illegible]reits 2004 formulierte Frage „*Is my B1 your B1?*“ Denn nicht jedes am *GeR* [illegible]gerichtete Modell von Sprachkompetenz ist deckungsgleich mit anderen Definitionen und nicht jede Verteilung von Schülerleistungen auf Skalen entspricht der in anderen Schulsystemen. Selbst wenn die jeweiligen Befunde statistisch untermauert sind, stellen sich viele Fragen, wie das methodisch erfolgt ist.

Hinzu kommt, dass der *GeR* bzw. der *CEFR* überarbeitet wird. Aufgrund dieser Überarbeitung muss eventuell auch das Spektrum der Referenzniveaus neu „geeicht“ werden, da sowohl bei den niedrigen Niveaus „vor A1“, A1 und A2, als auch auf C2 einiges ins Rutschen gekommen ist. Das ist wichtig für die nächsten Abschnitte in diesem Kapitel, die sich mit Vergleichs- und Abschlussarbeiten beschäftigen werden – ein Thema, mit dem sich Lehrkräfte verstärkt auseinandersetzen müssen.

[3] https://www.iqb.hu-berlin.de/bt (letzter Zugriff: 01.12.2017)

[4] Das Design und die Ergebnisse der beiden Erhebungen sind in einem Beitrag von Neil Jones in der Zeitschrift für Fremdsprachenforschung (ZFF 2016, 27/1, 39–57) nachzulesen.