Tema

NÜTZLICHE(RE) SPRACHTESTS

BEISPIELE FÜR DIE FORMATIVE BEURTEILUNGS-PRAXIS VOR DEM HINTERGRUND EINER NEUEREN PERSPEKTIVE AUF TEST-GÜTEKRITERIEN

Practicable? – Definitely! Reliable? – As much as possible. Valid? – Yes, at least to some degree! Practicable, reliable, valid: language assessment - from standardized language testing to informal classroom assessment - must be all of this, and even more. Language assessments need to suit their purpose and be *useful* to those concerned by them, to teachers and learners in particular. This is how recent conceptualizations of test quality can be summarized (cf. Green, 2014; Bachman & Palmer, 2010; Grotjahn & Kleppin, 2015). With this in mind, first the concept of usefulness will be presented. This is then used as a point of departure for a closer look at two very different forms of assessment: C-Tests and Dynamic Assessments. Both forms are still not commonly found in practice but can be very useful additions to a teacher's assessment toolkit.

Thomas Studer | Fribourg

Thomas Studer ist assoziierter Professor für Deutsch als Fremd- und Zweisprache an der Universität Freiburg/ Fribourg und Direktionsmitglied des Freiburger Instituts für Mehrsprachigkeit. Zu seinen Forschungs- und Arbeitsschwerpunkten gehören die Sprachlehr- und Sprachlernforschung, die Fremdsprachendidaktik und das Testen und Prüfen.

PIÙ ARTICOLI SU QUESTO TEMA:
WWW.BABYLONIA.CH >
ARCHIVIO TEMATICO > SCHEDA 6

Praktikabel? Unbedingt! Zuverlässig (reliabel)? Soweit möglich. Gültig (valide)? Ja, mindestens teilweise! – Praktikabel, zuverlässig, gültig: All diese Eigenschaften und noch einige mehr sollten Sprachtests aufweisen. Das gilt nicht nur für standardisierte Tests, sondern eingeschränkt auch für informelle Assessments im Unterricht. Es kommt aber noch etwas dazu: Über die Standard-Merkmale hinaus oder sogar in erster Linie sollten Sprachtests, gleich welchen Kalibers, das Kriterium der Nützlichkeit erfüllen. Tests sollten nützlich sein, und zwar für Lehrende *und* Lernende. So lässt sich eine neuere Perspektive auf Qualitäts-Kriterien für Tests umreissen (u.a. Green, 2014; Bachman & Palmer, 2010 und auch Grotjahn & Kleppin, 2015). Im Folgenden soll diese Nützlichkeits-Forderung zunächst dargestellt (Kap. 1) und dann zum Anlass genommen werden, exemplarisch zwei sehr verschiedene Tests näher zu beleuchten: Varianten des C-Tests einerseits (Kap. 2) und sog. dynamische Assessments andererseits (Kap. 3). Beide Verfahren sind noch wenig verbreitet, bringen aber ein beachtliches Nützlichkeits-Potenzial mit. Das macht sie für die schulische Praxis interessant, insbesondere für eine auf (Förder-)Diagnostik ausgerichtete Test- und Beurteilungspraxis.

1. Qualitäts-Kriterien für Tests

Green (2014) stellt die Qualitäts-Kriterien für einen Test in Form eines Kegels dar und bringt damit eine Wichtigkeits-Reihenfolge dieser Kriterien zum Ausdruck (Siehe Abb. 1).

Nützlichkeit als zentrales Qualitäts-Kriterium

An der Spitze der Pyramide stehen *nützliche Konsequenzen*, weil es sich dabei um das ultimative Ziel eines jeden Testsystems handelt (Green, ebd.: 59). Die Spitzenposition der „*beneficial consequences*" ist so zu verstehen, dass diesem Kriterium durch Berücksichtigung der anderen drei Kriterien zwar zugearbeitet werden kann und soll, dass sich nützliche Konsequenzen aber nicht automatisch aus diesen anderen Kriterien ergeben, sondern schon bei der Testauswahl resp. bereits zu Beginn der Testerstellung ‚offensiv' angegangen werden sollten. Um-

gekehrt werden nützliche Konsequenzen kaum zu erreichen sein, wenn es einem Test (weitgehend und gleichzeitig) an Praktikabilität, Reliabilität und Validität mangelt.

Unbestritten ist, dass ohne *Praktikabilität* – bei Bachman & Palmer (2010: 262) durch die Differenz zwischen den für die Testerstellung nötigen und den dafür vorhandenen Ressourcen bestimmt – gar nichts geht. Bei den hierarchisch höher stehenden Kriterien wird es dann komplexer und zunehmend kontrovers: Während es für die *Zuverlässigkeit* oder Verlässlichkeit von Messinstrumenten und Prüfungsergebnissen noch handhabbare Prüf-Mechanismen gibt (z.B. Zweitkorrektur einer Schreibaufgabe), sind mit *Validität* eine Reihe von schwierigen Fragen verbunden. Veranschaulicht anhand eines Hörtests aus dem Lehrbuch – etwa ein Interview, das mit einigen Richtig-Falsch-Items verbunden ist –, geht es z.B. um folgende Fragen: Kann ich mit diesem einen Test gültige Aussagen über die Hörverstehens-Kompetenz der Lernenden machen? Oder brauche ich dafür mehrere Texte mit unterschiedlichen Textfunktionen, die von verschiedenen SprecherInnen gesprochen werden, sowie eine Mindest-Anzahl von Items? Nach welchen Regeln müssen diese Items konstruiert sein, damit sie – einzeln und insgesamt – etwas Relevantes über die Hörverstehens-Kompetenz aussagen? Und wie umfangreich müssen die Vor-Tests sein, damit ich das alles feststellen oder zumindest besser abschätzen kann? Unter *nützlichen Konsequenzen* eines Tests schliesslich lässt sich ganz Verschiedenes subsummieren; primär wohl das, was andernorts *Impact*, *Washback* oder *Backwash* heisst: Gemeint ist die Tatsache, dass jeder Test Auswirkungen auf verschiedenen Ebenen hat, darunter ‚gut gemeinte', d.h. intendierte und in der Regel positive ebenso wie nicht-intendierte (meist negative). Positive Auswirkungen von Tests im schulischen Kontext (Mikro-Ebene) sind z.B. gezielte Hörverstehens-Übungen als Folge von grösseren Tests, in denen das Hörverstehen berücksichtigt oder aufgewertet wird, oder Flüssigkeits-Trainings als Folge von Tests zum Sprechen. Negative Wirkungen können sich u.a. einstellen, wenn durch eine hohe Kadenz summativer Tests Lerngegenstände zurückgedrängt werden, die nicht prüfungsrelevant sind[1].

Konsequenzen für die Testpraxis

Je weiter nach oben man also in Greens Kegel geht, desto komplexer und anspruchsvoller werden die Anforderungen an qualitativ gute Tests und desto kontroverser sind die Positionen. Doch was bedeutet das für die Testpraxis? Zweierlei: Zum einen wird es um ein Abwägen der Qualitäts-Kriterien gehen, seien es nun die vier Gross-Kriterien nach Green, die sechs Kriterien nach Bachman & Palmer oder die insgesamt elf Aspekte der Testqualität bei Grotjahn & Kleppin[2]. So wird man bei der Überprüfung des Hörverstehens Abstriche bei der Reliabilität in Kauf nehmen und keine aufwändigen Abklärungen im Bereich der Validität machen können. Unverzichtbar hingegen ist immer die zur Validität zählende Frage, „ob und in welchem Umfang die Aussagen und Entscheidungen, die wir anhand der beobachteten Leistungen der Prüflinge treffen, gerechtfertigt sind." (Grotjahn & Kleppin, 2015: 50) Zu den leicht realisierbaren Proben in diesem Bereich sollte mindestens die Überprüfung der folgenden beiden, vermeintlich einfachen Fragen gehören: Können die Items eines Hörverstehens-Tests auch ohne den Text gelöst werden? Sind die Items tatsächlich eindeutig lösbar und voneinander unabhängig, sodass Testteilnehmende, die ein Item falsch lösen, eine echte zweite Chance haben? Treten bei diesen Proben Mängel auf, sind gültige Aussagen über die Leistungen der ProbandInnen nicht möglich; der Test geht an der Hörverstehens-Kompetenz vorbei bzw. ist nicht valide.

Das gerade angeführte Zitat von Grotjahn & Kleppin verweist noch auf etwas anderes, nämlich auf eine Dynamisierung des Validitäts-Begriffs und damit auf eine Umorientierung in der Validitäts-Theorie, wie sie bereits Messick (1989: 13) gefordert hatte:

> *Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails.*

Nach Messick kommt es letztlich ganz zentral auf die Schlussfolgerungen an, die man auf Basis der Testresultate macht, nicht (so sehr) auf den Test selbst. Dieses Messick-Zitat lässt sich, das ist der zweite Aspekt, sehr wohl auch ganz praktisch

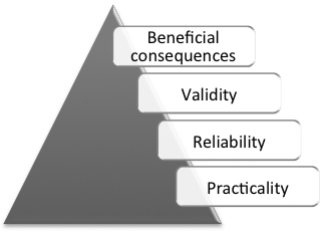

Abb. 1: Vier Qualitäts-Kriterien für Tests (Green, 2014: 58; hier als Pyramide reproduziert)

1 Wash-Back-Effekte werden manchmal auch als Aspekt von Validität angesehen (vgl. den Ausdruck *consequential validity*) und würden dann mit zu dem gehören, was einen gültigen Test ausmacht.

2 Im Unterschied zu Green (2014) sind bei Bachman & Palmer (2010) auch *Authentizität* (Inwieweit korrespondieren Merkmale der Testaufgabe mit Merkmalen von Aufgaben ausserhalb der Testsituation?) und *Interaktivität* (Sind die Testaufgaben so gestaltet, dass die Testteilnehmenden tatsächlich das zeigen, was sie können?) separate Kriterien. Grotjahn & Kleppin (2015) halten ausserdem am Objektivitäts-Kriterium fest und weisen zusätzlich Fairness und Transparenz sowie die (Item-bezogenen) Merkmale *Trennschärfe* und *Schwierigkeit/Leichtigkeit* separat aus. Im Wesentlichen sind alle drei Ansätze miteinander kompatibel, unterschiedlich weit gefasst wird v.a. das Validitäts-Kriterium. Fairness z.B. – u.a. umschrieben als Anspruch, niemanden z.B. aufgrund des Geschlechts oder durch ‚Tabuthemen' zu benachteiligen – ist zweifellos ein zentrales Güte-Kriterium, kann aber, ebenso wie der Back-Wash-Effekt, als Aspekt von Validität angesehen werden. Kurz gefasst ergibt sich die folgende Maxime: Als gültig erweist sich ein Test erst dann, wenn sich Fairness und intendierte Auswirkungen nachweisen lassen.

verstehen: ‚Magere' Tests (praktikable, aber unzuverlässige und bezüglich Gültigkeit nicht wenigstens minimal geprüfte) wird es immer geben, aber solche Tests sind keine Basis für Schlussfolgerungen auf die Kompetenz der Lernenden, und Anschlusshandlungen wie z.B. die Vergabe von Abschlussnoten sind auf dieser Grundlage nicht gerechtfertigt.

Im konzeptuellen Bereich schliesst sich mit Messick ein Kreis: Wenn Schlussfolgerungen ausschlaggebend sind und entscheidend ist, ob die Anschlusshandlungen in der Folge eines Tests gerechtfertigt sind, ist man wieder an der Spitze der Hierarchie von Test-Gütekriterien angelangt, wie sie Green vorschlägt (vgl. Abb. 1): Das Zentrale beim Testen sind die Konsequenzen, die ein Test haben sollte. Zusammenfassend kann man also sagen: Möglichst nützliche Konsequenzen sind das Leitkriterium für die Qualität von Tests im Dickicht der Tests-Gütekriterien.

Abb. 2: Beispiele für drei C-Test-Varianten, Exzerpte aus Baur et al. *(2013: 5, 14 und 15);*
1) C-Test zur Überprüfung der allgemeinen Sprachfähigkeit und Sprachrichtigkeit und des Textverständnisses
2) C-Test zur Überprüfung der Morphologie in Nominalphrasen
3) C-Test zur Überprüfung von Fachwortschatz

1) Steinzeitmenschen
Schon vor Tausenden von Jahren lebten Menschen auf der Erde. Sie benutzten Werkze_____ aus Stein. Di_____ Zeit heißt Steinze_____ . Die Steinzeitmenschen jag_____ oft kleine Ti_____ ...

2) Die Fernbedienung
Ich bin komplett durcheinander. Ich habe bei mir i_____ mei_____ Zimmer ferngeschaut und wollte v_____ d_____ ers_____ Programm, vom Tennis, auf das Zweite schalten. ...

3) Das Gebiss des Wolfes
Das Gebiss des Wolfes ist besonders gut zum Fleischfressen geeignet. Die dolchartigen _____kzähne dienen zum _____sthalten und Töten der _____te. ...

Wann ist ein Test nützlich?

Nützlich ist ein Test für Lehrende dann, wenn er z.B. dabei hilft, Lernziele zu überprüfen und ggf. anzupassen, eine einzelne Lektion zeitlich und/oder inhaltlich zu justieren oder Massnahmen der Binnendifferenzierung zu treffen. Lernende wiederum können *dann* Nutzen aus einem Test ziehen, wenn sie ein reichhaltiges, möglichst individuelles Feedback zum Testergebnis bekommen, das sie beim Weiterlernen unterstützt[3]. Der allgemeine Zweck des Testens und Prüfens besteht darin, Informationen über Lernstände und -entwicklungen zu beschaffen (vgl. z.B. Studer, 2012). Beim Testen und Prüfen *im Unterricht* kommt es zentral darauf an, nicht nur bilanzierende Informationen über Lernstände im Anschluss an eine längere Lernphase zu bekommen (*summative assessments*), sondern auch Informationen, die auf Lernprozesse und Lernfortschritte bezogen sind, denn solche Informationen kommen dem aktuellen Lerngeschehen näher und zielen auf den Entwicklungshorizont der Lernenden (*formative assessments*). Wichtig ist dabei, dass Informationen aus summativen Tests durchaus auch für formative Zwecke genutzt werden können (vgl. etwa den GER (2001: 81); Dunn & Mulvenon (2009) erwägen weitergehend auch die von ihnen so genannten „Evaluation" formativer Befunde für summative Zwecke).

Vor diesem Hintergrund sollen nun zwei noch weniger bekannte Test-Verfahren mit intakter Praktikabilität und gutem Nützlichkeits-Potenzial besprochen werden, die für formative Zwecke eingesetzt werden können: Varianten des C-Tests (Kap. 2) und dynamische Assessments (Kap. 3).

2. C-Tests: Überraschend vielfältig

Beim C-Test, einer besonderen Form von Lücken-Test („C" steht für *cloze*), gab es in letzter Zeit neuere Entwicklungen u.a. für Deutsch als Zweitsprache in der Schule (siehe Baur *et al.*, 2013). Diese Entwicklungen lassen sich auch für die Beurteilung im Fremdsprachenunterricht nutzen, und zwar etwa ab der sechsten Klasse (der achten Klasse nach HarmoS), weil für diese Tests eine basale Lese- und Schreibkompetenz in der Zielsprache Voraussetzung ist. Ausgehend vom klassischen C-Test[4] lassen sich zwei Varianten unterscheiden: a) Eine Variante, mit der sich einerseits die allgemeine Sprachfähigkeit und Sprachrichtigkeit und andererseits das Textverständnis überprüfen lässt und b) eine Variante zur Beurteilung von sprachlichen Teilfertigkeiten, namentlich von linguistisch-grammatischen Kompetenzen (z.B. Beherrschung von Präteritums-Formen) und von Fachwortschatz. Beide Varianten können von Lehrpersonen gemäss den eigenen Bedürfnissen und mit kleinerem Aufwand selbst erstellt und auch selbst eingesetzt und ausgewertet werden. Die folgende Darstellung ist auf die Beispiele in Abb. 2 bezogen und an Baur *et al.* (2013) orientiert.

3 Zu präzisieren ist, dass es sich bei diesem Feedback vorerst nur um ein Lern*angebot* handelt. Wie dieses Angebot von den Lehrpersonen arrangiert wird und ob und wie es dann von den Lernenden genutzt wird, sind andere, weitgehend offene Fragen (vgl. u.a., mit einiger Skepsis, Dunn & Mulvenon, 2009; optimistischer ist z.B. Davin, 2013, im Kontext des *dynamic assessment*, vgl. unten, Kap. 3).

4 Zum Forschungsstand mit besonderer Berücksichtigung der Validitäts-Diskussion vgl. u.a. Grotjahn (2011).

a) C-Tests zur Überprüfung der allgemeinen Sprachfähigkeit und Sprachrichtigkeit und des Textverständnisses

Bei C-Tests gemäss a) wird ein Text im Umfang von höchstens etwa 200 Wörtern nach der Methode einer 3er-Tilgung bearbeitet, d.h. in jedem dritten Wort wird die zweite Hälfte gelöscht und durch eine durchgehende Linie ersetzt (vgl. Abb. 2.1: DiESE, jagTEN, TiERE). Bei ungerader Buchstaben-Anzahl wird ein Buchstabe mehr getilgt. Komposita werden so bearbeitet, dass nur die Hälfte der letzten bedeutungstragenden Einheit gelöscht wird (z.B. Werk-zeUGE, Stein-zeIT in Abb. 2.1). Eingerahmt wird der so bearbeitete Text von je einem unveränderten Einleitungs- und Schlusssatz, damit das Thema des Textes trotz Lücken erfasst werden kann. Üblicherweise werden in dieser Art vier bis fünf thematisch verschiedene Texte à je 20 Lücken bearbeitet, sodass sich insgesamt 80 bis 100 Lücken bzw. Items ergeben. Die Texte sollten den SchülerInnen inhaltlich bekannt sein, kommen also z.B. aus dem Lehrbuch, können aber gegenüber dem Original gekürzt und leicht verändert (i.d.R. vereinfacht) werden. Pro Text wird ca. fünf Minuten Lösungszeit veranschlagt, sodass der gesamte Test in 20-30 Minuten durchführbar ist. Die Instruktion für die Lernenden ist einfach („ergänze die Lücken"; Hinweis auf zügiges Arbeitstempo und, während der Testdurchführung, auf noch verfügbare Zeit). Wichtig ist, dass das ungewohnte Verfahren anhand eines Beispiels geübt wird. Dabei sollte den Lernenden auch das Auswertungsverfahren erklärt werden (gilt insbesondere auch für den WE-Wert, s. den übernächsten Abschnitt).

Verglichen mit dem klassischen C-Test (Löschung der zweiten Hälfte von jedem zweiten Wort, Punkte für jeden fehlenden Buchstaben) bietet diese für SchülerInnen entwickelte Variante mehr Anhaltspunkte für die inhaltliche Texterschliessung, was technisch auch durch die Wahl von Linien anstelle von Punkten unterstützt wird; zu diesen und weiteren Einzelheiten vgl. Baur *et al*. (2013: 3ff.). Gegenüber dem klassischen C-Test unverändert bleibt dagegen das theoretische Konzept der reduzierten sprachlichen Redundanz als Grundidee dieser Test-Form.

Das Besondere der C-Testvariante a) besteht darin, dass bei der Auswertung zwei Werte ermittelt werden: Ein Richtig/Falsch-Wert (R/F-Wert) und ein Worterkennungswert (WE-Wert). Der R/F-Wert ergibt sich aus der Anzahl der vollständig richtig (d.h. semantisch UND grammatikalisch UND orthografisch korrekt) ergänzten Lücken, z.B. SteinzeIT; der WE-Wert ergibt sich aus der Anzahl der semantisch korrekt ergänzten Lücken, d.h. neben SteinzeIT wäre z.B. auch SteinzeIHT akzeptabel, nicht aber SteinzeBRA. R/F- und WE-Wert erlauben nützliche Einblicke in das individuelle Sprachvermögen der SchülerInnen, entlang folgender drei Interpretationsregeln (Baur *et al.,* ebd.: 9):

1. Keine oder eine geringe Differenz zwischen R/F- und WE-Wert im oberen Punktbereich sprechen für eine gute allgemeine Sprachkompetenz.
2. Eine eher kleine Differenz zwischen R/F- und WE-Wert im unteren Punktbereich lässt auf fehlendes Textverständnis schliessen.
3. Höhere WE- als R/F-Werte im und ab dem mittleren Punktbereich deuten auf intaktes Textverständnis und gleichzeitig auch auf Schwierigkeiten im formalsprachlichen Bereich hin.

b) C-Tests zur Überprüfung von Teilfertigkeiten

C-Tests zur Überprüfung von Teilfertigkeiten funktionieren im Prinzip gleich wie die unter a) besprochenen Tests, mit folgenden Modifikationen: Bei der Ermittlung *linguistisch-grammatischer Kompetenzen* wird zuerst diejenige zielsprachliche Form oder Struktur bestimmt, deren Beherrschung überprüft werden soll, z.B. Präteritums-Formen, Endungen in Nominal- und Präpositionalphrasen, Pronomina. In Abb. 2.2 ist die Morphologie in Nominalphrasen das Zielphänomen. Mit diesem Fokus wurde der Text an *den* Stellen bearbeitet, an denen das Zielphänomen realisiert ist (z.B. iN meiNEM Zimmer, vON dEM ersTEN Programm in Abb. 2.2). Bei dieser Variante des C-Tests reicht es in der Regel, nur etwa zehn Lücken pro Text zu bestimmen; wollte man mehr Lücken erreichen, müssten die Texte zu stark bearbeitet werden und würden unnatürlich. Unter diesen Bedingungen resultieren für einen gesamten Test 40 bis 50 Items (vier bis fünf Texte à je zehn Lücken).

Wichtig bei Teilfertigkeits-Tests zur Grammatik ist, dass jeweils nur *ein* Phänomen pro Gesamttest überprüft wird, ansonsten ist die Validität stark eingeschränkt. Werden mehrere Phänomene gleichzeitig überprüft, ist der Blick auf Stärken und Schwächen der SchülerInnen

Praktikabel, zuverlässig, gültig: All diese Eigenschaften und noch einige mehr sollten Sprachtests aufweisen. Das gilt nicht nur für standardisierte Tests, sondern eingeschränkt auch für informelle Assessments im Unterricht.

bei *einer* Zielsprachen-Struktur verstellt; an diesem Problem kranken viele selbst gemachte Lückentexte. Bei der Auswertung solcher Teilfertigkeitstests macht die Unterscheidung zwischen R/F-Wert und WE-Wert natürlich keinen Sinn: Da es um Grammatik geht, werden nur die (vollständig) richtigen Lösungen gezählt. Der R/F-Wert gibt recht differenziert Auskunft über die Beherrschung (genauer: Reproduzierbarkeit) der ausgewählten grammatischen Struktur oder Form.

Eine weitere Art von C-Tests überprüft den erarbeiteten *Fachwortschatz*, was insbesondere für den zweisprachigen Sachfachunterricht interessante Perspektiven bietet. Im Unterschied zu den bisher besprochenen Testvarianten wird beim Teilfertigkeitstest zum Fachwortschatz Stammtilgung der zu überprüfenden Wörter empfohlen. Abb. 2.3 zeigt den Anfang eines bearbeiteten Texts aus dem Fachbereich Biologie. Gelöscht und durch einen Strich ersetzt wurde dort jeweils die erste Hälfte der Zielwörter, z.B. ECk-zähne und BEute. Überprüfungen des Fachwortschatzes können pro Text zehn bis 20 Lücken umfassen, wobei die Textinhalte und Inhaltswörter den SchülerInnen bekannt sein müssen. Bei der Auswertung dieses Teilfertigkeitstests stehen wieder die R/F-Werte im Vordergrund.

Vorteile und Einschränkungen

Die drei vorgestellten C-Test-Varianten, insbesondere die Teilfertigkeits-Tests, können nicht nur als (formative und summative) Tests, sondern auch als Übungsformen, etwa als Gruppenarbeit, eingesetzt werden, um die SchülerInnen für den korrekten Sprachgebrauch zu sensibilisieren und die Aktivierung von Lese- und Lösungsstrategien zu fördern (Baur *et al*. 2013: 13, 15). Zusammengefasst sind die Vorteile dieser Tests folgende: gute Anpassbarkeit an die unterrichtete Zielgruppe (Schulstufe, Alter und Erfahrungshorizont der SchülerInnen) und spezifische Bedürfnisse (z.B. gerade behandelter Lernstoff), hohe Testökonomie bei der Konstruktion, Durchführung und Auswertung und somit gute Praktikabilität. Um eine befriedigende Validität zu erreichen, sollten selbst entwickelte C-Tests an erwachsenen Muttersprachlern erprobt werden. Können Muttersprachler die Tests nicht eindeutig und problemlos lösen, müssen die betroffenen Lücken revidiert werden. Geht der mit dem Test verbundene Anspruch über formative Zwecke hinaus in Richtung eines grösseren Abschlusstests, ist darüber hinaus eine minimale Erprobung in ein bis zwei Klassen unentbehrlich[5]. Hinsichtlich Nützlichkeit vermögen C-Tests der besprochenen Art zu überzeugen: Sie liefern je nach Schwerpunkt differenzierte Anhaltspunkte entweder zur (allgemeinen schriftsprachlichen) Sprachkompetenz und dem Textverständnis der SchülerInnen oder zu sprachlichen Teilkompetenzen, an denen im Unterricht gerade gearbeitet wird.

Zentral ist, dass C-Tests nicht als Ersatz für andere, z.B. kommunikative Verfahren, verwendet werden, sondern als Ergänzung dazu. So ist z.B. offen, ob das in einem C-Test unter Beweis gestellte Form- und Strukturwissen auch beim freien Schreiben verwendet wird, und generell sollten C-Tests nicht dazu (ver-) führen, den Schwerpunkt des Unterrichts doch wieder auf die Grammatik und schriftsprachliche Korrektheit zu legen (siehe dazu oben, Kap. 1, negativer Backwash).

3. Dynamische Assessments: Lernpotenziale im Blick

Dynamische Beurteilung (*dynamic assessment*, DA) spielt in der Pädagogik und besonders bei der Arbeit mit Lernenden, die so genannte ‚besondere Bedürfnisse' haben, seit langem eine Rolle, ist aber in der Fremdsprachenforschung ein relativ neues Phänomen. Zum fremdsprachenspezifischen DA gibt es inzwischen eine Reihe von theoretischen Auseinandersetzungen, erst wenige empirische Studien (mit kleiner Probandenzahl) und eine Vielfalt verschiedener Anwendungen, die von Lehr-Lern-Gesprächen mit unterschiedlichen Graden der (Vor-) Strukturierung über systematische Arten der Fehlerkorrektur beim Schreiben bis hin zu standardisierten, computerbasierten Tests zum Hören und Lesen reichen (Übersichten, an denen sich die folgende Darstellung hauptsächlich orientiert, bieten Grotjahn, 2015, und Kley, 2011)[6].

Im Unterschied zu den bekannten, oft statisch genannten Tests kommunikativer Sprachkompetenzen, wie sie etwa auch *lingualevel* bereit stellt (vgl. Lenz & Studer, 2007), zielt DA zwar auch auf den aktuellen Stand der Kompetenzentwicklung, strebt aber darüber hinaus und vor allem eine Diagnose des Lernpotenzials an und ist direkter auf die Förderung der Lernenden ausgerichtet. Das zentrale Element des DA sind Hilfestellungen,

5 Für belastbare förderdiagnostische Aussagen mit eigenen C-Test-Varianten bräuchte es Standardisierungen, in deren Rahmen auch Referenzwerte (Durchschnittswerte der Zielgruppen) ermittelt werden. Entsprechende Untersuchungen sind für einzelne Lehrpersonen natürlich nicht mehr praktikabel, sie sollten aber an die Hand genommen werden, wenn ein selbst entwickelter C-Test z.B. als Übertritts-Prüfung verwendet werden soll.

6 Wesentlich geprägt wird das Forschungsfeld DA durch die Arbeiten von Lantolf und Poehner, vgl. u.a. Poehner (2008), Lantolf & Pohener (2011). Eine praktisch ausgerichtete Darstellung ist Grotjahn & Kleppin (2015), Kap. 6.1.2; ein *computerized dynamic assessment* haben Poehner *et al*. (2015) entwickelt.

welche die Lernenden erhalten resp. in Anspruch nehmen können, wenn sie bei der Bearbeitung von Aufgaben ‚anstehen'. Beim Sprechen z.B. bestehen die Hilfestellungen aus spezifischen Feedbacks der Lehrperson, wenn die Lernenden Fehler machen, vgl. Abb. 3.1.

Erläuterungen zum Transkript: Die Lernerin Molly bildet die Frage ¿Cómo le gusta hacer en Buenos Aires?, braucht aber das falsche Fragewort (*cómo* statt *qué*). Die Lehrperson führt Molly mit vier Hinweisen zum korrekten Fragesatz. Die ersten drei Hinweise – eine Pause und ein skeptischer Blick (*prompt 1*), die Wiederholung der gesamten falschen Äusserung (*prompt 2*) und die Wiederholung des nicht korrekten Frageworts (*prompt 3*) – führen nicht zum Ziel. Erst der vierte Hinweis, eine Entscheidungsfrage (¿Cómo o qué?), kann von der Lernerin zur Verbesserung ihrer Äusserung genutzt werden.

Typischerweise sind die Rückmeldungen beim DA zunächst implizit (z.B. eine nonverbale, mimische Reaktion) und werden dann immer expliziter, wobei die Richtigstellung des Fehlers (der explizite Pol des Hinweis-Spektrums) nur dann erfolgt, wenn die Lernenden nicht vorher selbst reagieren. Die Idee ist also, dem/der Lernenden quantitativ und qualitativ genau *die* Hilfen anzubieten, die er oder sie braucht, um den Fehler zu erkennen und selbst zu korrigieren. Und genau diese Idee ist dann auch ausschlaggebend für die Bestimmung des Lernpotenzials: Die Indikatoren für das Lernpotenzial sind die Art und der Umfang der beanspruchten Hilfen. Beispielsweise ist die gerade gemachte Feststellung, dass die Lernerin Molly auf die impliziteren Hinweise der Lehrperson nicht reagiert (und relativ viel Hilfe beansprucht, bis die Selbstkorrektur gelingt), ein Indiz für ein noch eher kleines Lernpotenzial, was die behandelten WH-Fragen betrifft. Beides, Qualität und Quantität der beanspruchten Hilfen, kann sich von Lerner/in zu Lerner/in unterscheiden – auch, und das ist hervorzuheben, bei gleichem Kompetenz-Stand. Diese Perspektivenerweiterung, der Blick auf individuelle Lernpotenziale, ist der Mehrwert des DA gegenüber ‚statischen' Tests.

Das theoretische Fundament des DA sind die soziokulturelle Theorie Vygotskijs und dessen Konzept einer Zone der nächsten Entwicklung (*zone of proximal development*, ZPD), definiert als Differenz zwischen dem, was Lernende selbständig leisten können und dem, was sie zu leisten im Stande sind, wenn sie unterstützt werden[7].

1. **Molly:** ¿Cómo le gusta hacer en Buenos Aires? [How do you like to do in Buenos Aires?]
2. **Teacher:** (Prompt 1—paused and looked skeptically at student)
3. **Molly:** ¿Cómo . . . [How . . .]
4. **Teacher:** (Prompt 2) ¿Cómo le gusta hacer en Buenos Aires? [How do you like to do in Buenos Aires?]
5. **Molly:** ¿Dónde? [Where?]
6. **Teacher:** (Prompt 3) ¿Cómo? [How?]
7. **Molly:** (silence)
8. **Teacher:** (Prompt 4) ¿Cómo o qué? [How or what?]
9. **Molly:** ¿Qué le gusta hacer? [What do you like to do?]
10. **Teacher:** Sí, ¿Qué significa qué en ingles? [Yes, what does qué mean in English?]

Abb. 3.1: Classroom DA*: Einzelgespräch Lehrperson-Molly, während die Klasse zuhört; Kontext: Spanisch als Fremdsprache für Kinder, ca. A2-Niveau, Thema* WH-question formulation*; Exzerpt aus Davin (2013: 16).*

Interventionistische und interaktionistische Ansätze

Die Vielfalt von DA-Konzepten lässt sich grob zweiteilen in interventionistische und interaktionistische Ansätze (Grotjahn, 2015: 472f.; Kley, 2011: 68f.): Der *interventionistische Ansatz* folgt einem *Prätest-Intervention-Posttest-Design* und wird bisher v.a. in der Forschung praktiziert. Es bieten sich aber im Rahmen der Handlungs- und Aktionsforschung auch Umsetzungsmöglichkeiten im Unterricht an. Beim interventionistischen Ansatz geht es darum, die Lernstände zu Beginn der Untersuchung und nach der Intervention zu messen, und zwar mit gleichartigen, in der Regel ‚statischen' Tests und mit der Erwartung, dass sich die Leistungen beim zweiten Mess-Zeitpunkt verbessert haben[8]. Die Intervention selbst erfolgt nach den Prinzipien des DA, kann mehr oder weniger stark strukturiert sein und sich im Rahmen des Klassengesprächs auf jeweils eine/n Lernende/n beziehen oder in Gruppen realisiert werden. Abb. 3.1 ist ein Beispiel eines stark strukturierten Lehrer-Schüler-Gesprächs, in dessen Verlauf ein definiertes Set von Hinweisen eingesetzt wird. Davin (2013: 10) weist das folgende Hinweis-Set aus, das aus fünf nach Explizitheit geordneten Hinweisen besteht (Abb. 3.2)[9].

7 Genaueres und Weiteres zum theoretischen Fundament des DA u.a. bei Grotjahn (2015: 471f.). Den Grundsatz, dass Lernende mit Hilfen mehr und/oder anderes können als ohne Hilfen, kennt man auch aus den Can-do-Checklisten des Europäischen Sprachenportfolios.

8 Weiter ausgebaute Forschungs-Designs würden zusätzlich eine Kontrollgruppe einbeziehen und einen dritten Mess-Zeitpunkt ansetzen, um die Nachhaltigkeit des erwarteten Kompetenz-Zuwachses zu prüfen.

9 Ein illustratives Beispiel für den interventionistischen Ansatz des DA, bei dem in der Interventionsphase sowohl mit Lehrer-Schüler-Dialogen als auch mit Gruppen-Gesprächen gearbeitet wird, ist Davin (2013; 2011).

Abb. 3.2: Festgelegte Mediations-Hinweise der Lehrperson im Rahmen eines interventionistischen Ansatzes des DA (Davin 2013: 10)

Level of Explicitness	Mediation Prompt
Prompt 1	Pause with skeptical look
Prompt 2	Repetition of entire phrase by teacher with emphasis on location of error
Prompt 3	Repetition of specific site of error
Prompt 4	Forced choice option (i.e., when or where?)
Prompt 5	Correct formulation of question accompanied by explanation

Beim *interaktionistischen Ansatz*, einem sog. *Train- Within-Assessment-Design*, erfolgt grundsätzlich auf jedes „Item" eine Rückmeldung an die Lernenden. Beim Mündlichen geschieht dies in der Art der geschilderten Feedbacks bei Fehlern, wobei die Rückmeldungen weit vielfältiger und individueller sein können als in Abb. 3.2 dargestellt. Bei Tests zu den rezeptiven Fertigkeiten, z.B. bei Hörverstehens-Aufgaben mit Auswahl-Antworten, erfolgen die Rückmeldungen in Form von immer konkreteren Lösungshinweisen, wenn ein Item zuerst falsch gelöst wurde[10].

Bereits deutlich geworden sein sollte, dass insbesondere beim interaktionistischen DA Testen, Lehren und Lernen eng miteinander verzahnt sind. Testen und Unterrichten, sonst ja oft als Gegensatz dargestellt, werden hier zusammengeführt.

Deutlich geworden sein sollte weiter auch, dass den Lehrpersonen bei dynamischen Beurteilungen eine aktive und ganz entscheidende Rolle zukommt. Dies schlägt sich auch in der Begrifflichkeit nieder: Im Rahmen des DA sind Lehrpersonen nicht einfach Feedback-Lieferanten, sondern *Mediatoren*, womit gut zum Ausdruck kommt, dass es um die Vermittlung zwischen zielsprachlichen Strukturen/Aufgaben und den Lernenden geht, nicht bloss um Instruktionen oder Korrekturen der bekannten Art.

Aspekte der Auswertung

Was die Auswertung des DA angeht, eignen sich Beobachtungsprotokolle, und zwar auch solche für die Selbstbeobachtung der Lehrpersonen. Lehrerseitig wird es z.B. nützlich sein, sich anhand von Protokollen der eigenen Hilfestellungen und der Reaktionen der Lernenden darauf bewusst zu werden, sodass das Repertoire an Hilfestellungen ggf. angepasst werden kann (vgl. Abb. 3.3; Grotjahn & Kleppin, 2015: 134) führen für den interaktionistischen Ansatz des DA zehn verschiedene Typen von Hilfen an, die die Selbstkorrektur der SchülerInnen unterstützen).

Werden stärker strukturierte Formen des DA gemäss Abb. 3.2 eingesetzt, bieten sich zur Diagnose des Lernpotenzials der Lernenden Tabellen an, in die man für die jeweils beobachteten SchülerInnen Anzahl und Art der in Anspruch genommenen Hinweise einträgt. Idealerweise werden solche Eintragungen – beim gleichen Lernstoff! – mehrmals, mind. aber zu zwei Zeitpunkten gemacht, sodass sich auch Veränderungen des Lernpotenzials abschätzen lassen.

Vorteile und Einschränkungen

Die enge Verzahnung von Unterrichten und Testen bei dynamischen Assessments hat auch Folgen für die Interpretation der Test-Gütekriterien. Für den interaktionistischen Ansatz des DA lässt sich in etwa festhalten: Hinsichtlich Praktikabilität erweist sich das Verfahren einerseits als sehr flexibel, was Individualisierungs-Möglichkeiten und Lerngegenstände angeht, andererseits aber auch als recht aufwändig in Bezug auf Planung, Organisation und nachbereitende Reflexion. Validität ist dann gegeben, wenn das Ziel des DA erreicht wird, d.h. wenn es gelingt, das Lernpotenzial der Lernenden zu erfassen und die (weiteren) Lernprozesse zu unterstützen. *Wie* dies gelingt, d.h. mit welchen und wie vielen Hinweisen man die nächste Entwicklungszone ausleuchtet, kann von Lerner/in zu Lerner/in sehr verschieden sein und wird gerade deshalb kaum je eins zu eins wiederholbar sein, sodass Reliabilität bei diesem Verfahren kein sinnvolles Kriterium ist. Das Nützlichkeits-Potenzial schliesslich ist als gross einzuschätzen: Kenntnisse von „Leistungsreserven" (Grotjahn, 2015: 470) sind nicht nur für Lehrpersonen relevant (u.a. Hilfe für das Abstimmen des Unterrichts auf individuelle Kompetenzen und für das Setzen realistischer Lernziele), sondern nützen auch den SchülerInnen (Bewusstmachung von Lernvorgängen und Lernstrategien, auch als zentraler Bedingung für autonomes Lernen).

Kritisch anzumerken ist, dass dynamische Assessments zwar für eine breite Palette von Lerngegenstände in Frage kommen, aber erfassen und weiter entwickeln lassen sich doch nur Fähigkeiten

10 Die Lösungshinweise bei schriftlichen Tests können beim Durchsprechen der Aufgaben mündlich gegeben werden, es sind aber auch schriftliche und sogar standardisierte Rückmeldungen möglich, wie das im von Poehner *et al.* (2015) entwickelten Computer-basierten DA zum Hörverstehen realisiert ist.

Abb. 3.3: Protokoll zur Selbstbeobachtung der (Wirksamkeit der) Unterstützung von Lernenden beim interaktionistischen DA (Grotjahn & Kleppin, 2015: 136)

Hilfen zur Selbstkorrektur, die ich ausprobiert habe	Reaktionen der Lernenden, die Hinweise auf schon vorhandenes Potenzial geben

und Fertigkeiten, die dem Bewusstsein zugänglich sind, z.B. Strategie-Kenntnisse oder grammatisches Regelwissen. Nicht erfassen lassen sich dagegen unbewusste kognitive Prozesse, und unberücksichtigt bleibt der gesamte Bereich des impliziten Lernens. Klar ist schliesslich auch, dass sich der (erwartete) Erfolg des DA nicht über Nacht einstellen wird. Gerade der sehr flexible interaktionistische Ansatz des DA muss ausprobiert, reflektiert und trainiert werden, insbesondere auch, was den Umgang mit dem Hinweis-Repertoire angeht. Das aber – der Weg hin zu reflektierteren Beurteilungspraktiken – kann sich lohnen. Auf diesem Weg können mehr Sicherheit in der eigenen Beurteilungspraxis und mehr Klarheit über die Nützlichkeit der eingesetzten Verfahren erreicht werden.

Fazit

Jeder Test, auch ein unterrichtsnaher, kann und sollte noch vermehrt unter dem Aspekt beurteilt werden, ob er den beteiligten Lehrenden und Lernenden etwas nützt und worin genau der Nutzen besteht. Mit dieser Frage-Haltung lässt sich nicht zuletzt ein Auswahl- und Analysegesichtspunkt für bestehende Tests gewinnen *und* man bekommt eine Richtschnur für das Erstellen eigener Tests in die Hand. Bei den beiden hier diskutierten Tests, Varianten des C-Tests und dynamischen Assessments, handelt es sich um praktikable Verfahren mit beachtlichem Nützlichkeits-Potenzial. Das privilegiert sie für Try-Outs im eigenen Unterricht – als Ergänzung zu den üblichen Tests im Rahmen eines erweiterten Konzepts von Beurteilung.

Literatur

Bachman, L. & Palmer, D. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Baur, R.S., Goggin, M. & Wrede-Jackes, J. (2013). *Der C-Test: Einsatzmöglichkeiten im Bereich DaZ*. Universität Duisburg-Essen: pro-DaZ. Online (1.4.16): www.uni-due.de/imperia/md/content/prodaz/c_test_einsatzmoeglichkeiten_daz.pdf

Davin, K. J. & Donato, R. (2013). Student Collaboration and Teacher-Directed Classroom Dynamic Assessment: A Complementary Pairing. *Foreign Language Annals* 46/1, 5-22.

Davin, K. J. (2011). *Group dynamic assessment in an early foreign language learning program: Tracking movement through the zone of proximal development* [Unpublished doctoral dissertation]. University of Pittsburgh. Online (1.4.16): http://d-scholarship.pitt.edu/7269/1/DAVINKJ_ETD.pdf

Dunn, K. E. & Mulvenon, S. W. (2009). A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education. *Practical Assessment, Research & Evaluation* 14/7, 1-11.

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin u. a.: Langenscheidt.

Green, A. (2014). *Exploring language Assessment and Testing*. New York: Routledge.

Grotjahn, R. (2015). Dynamisches Assessment: Grundlagen, Probleme, Potenzial. In: J. Bäcker & A. Stauch (Hrsg.), *Konzepte aus der Sprachlehrforschung - Impulse für die Praxis*. Frankfurt am Main: Lang, pp. 469-488.

Grotjahn, R. (2011). C-Tests – Aspekte der Validität. *Deutsch als Fremdsprache* 48/3, 131-137.

Grotjahn, R. & Kleppin, K. (2015). *Prüfen, Testen, Evaluieren*. München: Klett-Langenscheidt.

Kley, K. (2011). Dynamic Assessment. Zusammenführung von Unterricht und Leistungsmessung. *Deutsch als Fremdsprache* 48/2, 67-74.

Lenz, P. & Studer, T. (2007). *lingualevel. Instrumente zur Evaluation von Fremdsprachenkompetenzen*. Bern: Schulverlag blmv.

Lantolf, J. P. & Poehner, M. E. (2011). *Dynamic assessment in the foreign language classroom: a teacher's guide, second edition*. The Pennsylvania State University: CALPER Publications.

Messick, S. (1989). Validity. In: R. L. Linn (Ed.), *Educational measurement*, 3rd ed. New York: American Council on Education and Macmillan, pp. 13-103.

Poehner, M. E. (2008). *Dynamic Assessment: A Vygotskyan Approach to Understanding and Promoting L2 Development*. Berlin: Springer.

Poehner, M. E., Zhang, J. & Lu, X. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing* 32/3, 337–357.

Studer, T. (2012). Leistungsbeurteilung: Testfunktionen als Orientierungshilfe. Positiver Washback: Mit lingualevel Sprachkompetenzen beurteilen. *Grundschulmagazin Praxis* 5, 7-12.