Tema

DEFINITION UND MESSUNG VON BILDUNGSSTANDARDS IM BEREICH FREMDSPRACHEN

The development, operationalisation and evaluation of educational standards has become an important task for educational scientists and linguistic researchers in any modern educational system. This article outlines the necessary steps, and attendant pitfalls, of such a process by taking the development work performed in Switzerland in the context of foreign language education as a point of reference. It argues that in order to fulfil their function, educational standards need to be backed by a well-founded theoretical model of the latent variable at stake. The next step is a rigorously controlled process of item development coupled with a representative validation study. Finally, the article shows that without rich and varied empirical data from a representative sample, discussing minimal or maximum standards is a fruitless exercise.

Stefan D. Keller | FHNW

Stefan D. Keller leitet die Professur Englische Didaktik und ihre Disziplinen an der Pädagogischen Hochschule FHNW und ist stellvertretender Direktor des Instituts für Bildungswissenschaften der Universität Basel. In seiner Forschung beschäftigt er sich mit der Konzeptualisierung, Förderung und Messung komplexer argumentativer Schreibkompetenzen in Englisch auf der Oberstufe. Seine Forschungsgruppe entwickelt auch die Testaufgaben für die „Checks" im Bildungsraum Nordwestschweiz (Englisch) mit.

Am 19. August 1998 stellten Vertreter der EDK das „Gesamtsprachenkonzept" vor, ein Bündel von 15 ausführlich begründeten Empfehlungen zur Frage, welche Fremdsprachen wann, wie lange, wie und mit was für Anforderungen unterrichtet werden sollen. Dabei wurde auch explizit eingefordert, dass die Kantone „die Transparenz und Kohärenz des Fremdsprachenlernens gesamtschweizerisch dadurch [gewährleisten sollen], dass sie für das Ende der obligatorischen Schulzeit **verbindliche Richtziele** vereinbaren" (EDK, 1998). Dieser Satz stand am Anfang eines Prozesses der **Definition und Messung von Bildungsstandards in den Fremdsprachen**, der im folgenden Beitrag kurz beleuchtet werden soll. Dabei wird aufgezeigt, welche wissenschaftlichen **Herausforderungen** in diesem Prozess zu bewältigen sind, wie diese in den letzten 10 Jahren angegangen wurden und welche Arbeiten noch ausstehen. Dabei wird im ersten Teil des Beitrags das Verhältnis von Unterrichtskultur und Bildungssteuerung thematisiert; im zweiten Teil stehen dann Entwicklungs- und Validierungsprozesse standardisierter Testverfahren im Zentrum.

1. Kompetenzorientierung zwischen Lernkultur und Testentwicklung

Die 1998 eingeforderten nationalen Bildungsziele wurden im Rahmen des HarmoS Projekts entwickelt und im Juni 2011 vorgestellt (EDK, 2011). Die Bildungsstandards orientierten sich stark am *Gemeinsamen Europäischen Referenzrahmen für Sprachen* (GER; Europarat, 2001). Der GER beschreibt ausführlich, was Lernende leisten müssen, „um eine Sprache für kommunikative Zwecke zu benutzen, und welche Kenntnisse und Fertigkeiten sie entwickeln müssen, um in der Lage zu sein, kommunikativ erfolgreich zu handeln" (Europarat, 2001: 14). Generell sollen Schülerinnen und Schüler am Ende der obligatorischen Schule in der ersten und zweiten Fremdsprache Niveau A2.2 in den Kompetenzbereichen Leseverstehen, Hörverstehen und Sprechen erreichen, für Schreiben gilt für beide Fremdsprachen Niveau A2.1 als Mindeststandard.

Die jüngste Entwicklung in diesem Prozess **einer zunehmenden Kompetenzorientierung** von Unterricht und Unter-

richtssteuerung ist die momentan laufende Einführung des „Lehrplans 21" (LP21). Der LP21 baut auf den HarmoS Standards auf und nimmt diese als Grundlage für die Basiskompetenzen, die jeweils am Ende eines Zyklus erreicht werden müssen (EDK, 2016). Wie die HarmoS Standards und der GER beschreibt der LP21 Kompetenzen im Sinne von Can-Do Statements und nennt *en détail* „fachliche, personale, soziale und methodische Kompetenzen, die die Schülerinnen und Schülern in den Fachbereichen erwerben sollen" (EDK, 2016, o.S.). Bildungsstandards sind demnach als Bündel von Kompetenzen zu verstehen, die so konkret beschrieben werden, dass sie in Aufgabenstellungen umgesetzt und prinzipiell mit Hilfe von Testverfahren erfasst werden können (Klieme, Avenarius, Blum *et al*., 2003).

Im psychometrischen Konzept von Kompetenz, welches sich an den Arbeiten von Weinert (2001) orientiert, versteht man als Kompetenzen nicht das Verhalten selbst, sondern die bei den Schülerinnen und Schüler **verfügbaren Fähigkeiten und Fertigkeiten**, die notwendig sind, um **bestimmte Probleme bzw. Aufgaben lösen** zu können. Kompetenzen sind also hypothetische Konstrukte bzw. latente Merkmale, welche erst in einem Prozess der **Operationalisierung** mit Hilfe von Messinstrumenten der Beobachtung zugänglich gemacht werden können (Köller, 2008). Solche Messinstrumente erlauben es dann, Annahmen über die Eigenschaften der Kompetenzen im Sinne von latenten Variablen zu treffen.

Auffällig beim LP21 ist die Mehrdimensionalität der dort definierten Kompetenzaspekte, welche als Kern des schulischen Bildungsauftrags spezifiziert ist. Damit die Lernenden zentrale Kompetenzstandards in einem Fach erfüllen, sollen sie…

› zentrale **fachliche Begriffe und Zusammenhänge** verstehen, sprachlich zum Ausdruck bringen und in Aufgabenzusammenhängen nutzen können;
› über fachbedeutsame (wahrnehmungs-, verständnis- oder urteilsbezogene, gestalterische, ästhetische, technische …) Fähigkeiten und Fertigkeiten zum **Lösen von Problemen** und zur Bewältigung von Aufgaben verfügen;
› auf vorhandenes Wissen zurückgreifen bzw. sich das **notwendige Wissen beschaffen**;
› ihr sachbezogenes Tun **zielorientiert planen** und in der Durchführung angemessene Handlungsentscheidungen treffen;
› Lerngelegenheiten aktiv und selbstmotiviert nutzen und dabei **Lernstrategien** einsetzen
› fähig sein, ihre Kompetenzen auch in **Zusammenarbeit mit anderen** einzusetzen (EDK, 2016, o.S.).

Das hier implizierte Konzept von Kompetenz geht deutlich über kognitive Wissensaspekte (Faktenwissen) hinaus und schliesst komplexe Urteils- und Bewertungsprozesse in authentischen Situationen mit ein; darüber hinaus motivationale, emotionale und volitionale Aspekte des Lernens, die Verwendung von Lernstrategien und die Ausbildung von Sozialkompetenzen.

Die Einführung von Kompetenzzielen wie jenen des LP21 setzen also entsprechende Entwicklungen in der fachlichen Lehr-Lernkultur voraus, weil die Qualität der Kompetenzen entscheidend von der **Qualität der Prozesse** abhängt, in denen diese erworben wurden (Keller, 2013b). Und Kompetenzziele können nur dann zu Veränderungen im alltäglichen Unterricht führen, wenn es gelingt, den Lehrkräften Wege aufzuzeigen, wie sie mit diesen Bildungszielen unterrichtlich konkret vorgehen können. Dazu gehört in erster Linie die Entwicklung von Lehrmitteln, Aufgabenszenarien, welche auf die besagten Kompetenzen bezogen sind und diese für die Lernenden konkret werden lassen. Darüber hinaus sollten Lehrkräfte lernen, diese Kompetenzvorgaben selber in erwartbares Können von Schülerinnen und Schülern und damit auch in konkrete Lernaufgaben zu übersetzen. Anstatt Wissen zu vermitteln und zu präsentieren müssen sie verstärkt Situationen schaffen, in denen sich Kompetenzen der Lernenden, oder Teile davon, äussern können. Nimmt man Aspekte wie Planung des eigenen Lernens und Einsatz von Lernstrategien wirklich ernst, so wird deutlich, dass jungen Menschen nicht einfach komplexitätsbereinigte Aufgaben vorgelegt werden dürfen (Keller, 2013b). Gleichzeitig können Kompetenzziele im Sinne von Bildungsstandards ihre Orientierungsfunktion nur erfüllen, wenn sie in Testverfahren umgesetzt und die zur Überprüfung notwendigen Instrumente grossflächig und ökonomisch eingesetzt werden können.

Zwischen diesen beiden Ansprüchen der Kompetenzorientierung liegt ein Widerspruch, der sich nur teilweise auflösen lässt: Einerseits geht es um die Entwicklung von Lehr-Lernformen, bei denen ne-

Solche Kompetenzmodelle für schulische Fachdomänen zu entwickeln ist eine genuine, meistens nur kooperativ zu bewältigende Aufgabe der Fachdidaktik und Erziehungswissenschaft/Psychologie, die der Definition von Standards zeitlich vorgeordnet sein muss.

ben kognitiven auch motivationale, ethische und soziale Komponenten der Leistung gefördert und ausgewiesen werden. Andererseits ist es sehr schwierig, viele dieser komplexen und weiterführenden **Kompetenzdimensionen in valide Testinstrumente** zu übersetzen. Bereits am Beispiel der PISA-Publikation von 2007 wurde illustriert, dass eine breite Fassung des Kompetenzkonstrukts unter Berücksichtigung motivationaler, volitionaler und selbstregulativer Aspekte nur in der Entwicklung getrennter Instrumente für die verschiedenen Aspekte münden kann (PISA, 2007). Entsprechende Arbeiten für urteilsbezogene oder ästhetische Kompetenzen, für motivationale oder volitionale Aspekte von Kompetenz(erwerb) oder für Lernstrategien sind für den Fremdsprachenunterricht in der Schweiz nicht ansatzweise umgesetzt. Sie werden zwar manchmal im Rahmen von Schulzeugnissen, personalisierten Rückmeldungen usw. beschrieben, fliessen jedoch ins nationale Bildungsmonitoring nicht ein und entfalten damit auch nur wenig Wirkung als Qualitätskriterien einer outputorientierten Bildungssteuerung. Für die didaktische Weiterentwicklung des Englischunterrichts ist es wichtig zu betonen, dass das Spektrum an angezielten Kompetenzen im Fremdsprachenunterricht breiter und komplexer ist als der relative schmale Ausschnitt davon, der im Rahmen von standardisierten Testverfahren operationalisiert werden kann.

2. Qualitätsaspekte bei der standardisierten Messung von Fremdsprachkompetenzen

Die momentan vorliegenden Standards, die tatsächlich in Testaufgaben operationalisiert wurden, betreffen die **vier klassischen *language skills*** d.h. Lese- und Hörverständnis, Sprechen und Schreiben. Sowohl bei der Validierung von Messinstrumenten für die nationalen Bildungsziele im Rahmen von HarmoS (2007), wie auch in Kantonalen Messungen wie den „Checks" im Bildungsraum Nordwestschweiz (Ender, Moser *et al*., 2017), wurden sowohl rezeptive wie auch produktive Sprachkompetenzen in den Blick genommen. Dabei sind unterschiedliche Testverfahren relevant, die im Folgenden beschrieben werden. Bei rezeptiven Kompetenzen (Hören und Lesen) erfolgt die Messung meist mit Aufgabenformaten wie *multiple choice* oder Lückentexten, wobei meist von „discrete item testing" gesprochen wird (Weigle, 2002). Bei der Messung von sprachproduktiven Leistungen wird „performance assessment" angewendet, wobei komplexe und authentische sprachliche Äusserungen (Texte, Dialoge, Vorträge, usw.) anhand von genau definierten Beurteilungskriterien (Rastern) durch menschliche Rater beurteilt werden.

In beiden Fällen benötigt man erstens ein theoretisch gut spezifiziertes **Modell** der entsprechenden Kompetenz, und zweitens eine Umsetzung in konkrete Situationen (*performance assessment*) oder Testitems (*discrete item testing*). Erst dann, wenn das hinter den Leistungen liegende Kompetenzkonstrukt mit all seinen Facetten, Interaktionen und Antezedenzen präzise beschrieben ist, wird es möglich, ein tragfähiges Messmodell zu spezifizieren, auf dessen Basis die Entwicklung von Items zur Messung der Kompetenzen gelingen kann (Köller, 2008). Zu einem fundierten Kompetenzmodell gehört notwendigerweise ein **nomologisches Netzwerk**, das die Beziehung der jeweiligen Kompetenz zu anderen Konstrukten spezifiziert. Nomologische Netzwerke helfen zu klären, ob die im jeweiligen Fachkontext entwickelten Standards überhaupt domänenspezifisch sind (Köller, 2008). Solche Kompetenzmodelle für schulische Fachdomänen zu entwickeln ist **eine genuine, meistens nur kooperativ zu bewältigende Aufgabe der Fachdidaktik und Erziehungswissenschaft/Psychologie, die der Definition von Standards zeitlich vorgeordnet sein muss**. Erst wenn man elaboriert hat, was beispielsweise unter Lesekompetenz zu verstehen ist, wird man aus der Konstruktdefinition Verhaltensweisen ableiten können, die bei hohen oder geringen Ausprägungen auf dem Konstrukt beobachtbar sein sollten. Die Entwicklung entsprechender Kompetenzmodelle und deren Umsetzung in Testinstrumente für den Fremdsprachenunterricht oblag in der Schweiz dem sog. HarmoS Konsortium Fremdsprachen; entsprechende Arbeiten wurden bereits 2007 publiziert (HarmoS, 2008). Anhand des Leseverstehens wird nun der Prozess aufgezeigt, wie auf der Basis gut elaborierter Kompetenzmodelle valide Testitems entwickelt werden können (*discrete item testing*).

Das Schweizer HarmoS Konsortium hatte den Vorteil, dass es sich bei seinen Entwicklungsarbeiten bereits auf Kompetenzmodelle beziehen konnte, welche im GER sowie in verschiedenen, darauf aufbauenden Arbeiten ausführlich spezifiziert und zudem in wissenschaftlichen und schulpraktischen Kreisen eine recht grosse Akzeptanz geniessen (HarmoS, 2008: 5). Zuerst wurden zwei **Arten des Lesens** spezifiziert: sorgfältig-genaues Lesen sowie erkundendes und selektives Lesen. In enger Anlehnung an PISA wurden zur weiteren Differenzierung des Konstrukts drei **Handlungsaspekte** unterschieden: Informationen entnehmen, Interpretieren und (sich) in Beziehung setzen oder auch „Evaluieren" (OECD, 2003). Da die Lesekompetenz der Lernenden auch von ihrem Textsortenwissen beeinflusst wird, wurden fünf unterschiedliche Texttypen berücksichtigt: deskriptiv/expositorisch narrativ, argumentativ und instruktiv. Hinzu kam die Gruppe der diskontinuierlichen Texte. z.B. Websites. Unterscheidungskriterien waren zentrale sprachliche Merkmale der Texte wie etwa häufig vorkommende Satzarten oder typisches Vokabular.

Insgesamt umfasste das Konstrukt fremdsprachliche Lesekompetenz also die folgenden Komponenten: Zwei Leseprozesse oder -arten, drei Handlungsaspekte des Lesens sowie fünf Texttypen. Damit wurde an die internationalen Arbeiten in diesem Bereich angeknüpft und Items entsprechend dieser Konzeption generiert (Alderson, 2000). Durch die klare Konstruktfassung und die internationale Anbindung hatten die Items ein klares theoretisches Fundament im Sinne eines nomologischen Netzwerks (vgl. oben).

Als Testinstrumente werden zur Überprüfung der einzelnen Kompetenzbereiche ausschliesslich **Kommunikationsaufgaben** (*communication tasks*) eingesetzt. Damit kann ein Bezug geschaffen werden zu plausiblen Situationen ausserhalb des Tests, in denen eine solche oder ähnliche Anforderungen möglicherweise bewältigt werden müssen. Zur genaueren Beschreibung der Items (z.B. Fragen, welche zu einem Text gestellt werden) wurden sog. *item maps* erstellt. Diese bestehen aus Merkmalen, von denen man annimmt, dass sie die Schwierigkeit der Items beeinflussen. Die spezifischen Ausprägungen dieser Merkmale können kodiert und mittels statistischer Verfahren mit der empirisch ermittelten Schwierigkeit der Items in Beziehung gesetzt werden. Solche Verfahren sollen dazu beitragen, die empirische Schwierigkeit von Items besser interpretieren und, darauf aufbauend, in Zukunft besser voraussagen zu können. In der Untersuchung zum Leseverstehen wurden insgesamt 11 potentiell schwierigkeitsbestimmende Merkmale ausgewählt, darunter die drei fokussierten Komponenten des Konstrukts fremdsprachliche Lesekompetenz, das Antwortformat der Testaufgaben sowie sieben weitere Merkmale. Diese Merkmale charakterisieren die Sprache der Lesetexte und besonders der Textpassagen genauer, auf die sich die Items beziehen. Damit werden auch ausgewählte linguistische Kompetenzen genauer beleuchtet, auf die zugegriffen werden muss, um die Aufgabenstellungen zu bearbeiten und die Texte korrekt zu verstehen: Anzahl Wörter, lexikalische Dichte des Textes, Lesbarkeit, Wortfrequenz, Textmenge, Grammatik und Lexik (HarmoS, 2008: 31-32).

In einer nationalen Validierungsuntersuchung wurden anschliessen total 47 Aufgaben zum Leseverstehen auf 32 verschiedene Testhefte verteilt. Jedes Heft bestand aus zwei bis drei Aufgaben und wurde von 150 bis 250 Schülerinnen und Schülern bearbeitet. Mittels dieser Validierung konnte eine einheitliche Skala für das Leseverstehen in Deutsch, Französisch und Englisch als Fremdsprachen konstruiert werden, wobei die gesamte Lernentwicklung von Jugendlichen vom 6. bis ins 9. Schuljahr auf einer einzigen, kontinuierlichen Skala verortet werden kann. Nur dank solchen gemeinsamen Skalen ist es möglich, längsschnittliche Studien der Kompetenzentwicklung durchzuführen oder Testergebnisse über die beiden beteiligten Landesteile und alle drei Fremdsprachen hinweg miteinander zu vergleichen.

Im Falle des Leseverstehens umfasst die Skala knapp 200 Items (aus rund 50 Testaufgaben), die alle einem Feinniveau des GER (Referenzskala Leseverständnis) zugeordnet sind und die durch Merkmale genauer beschrieben sind, welche sich auf das Kompetenzmodell Fremdsprachen (und spezifischer auf das Konstrukt fremdsprachliche Lesekompetenz) beziehen.

Die Arbeiten des Harmos Konsortiums zeigen den Aufwand, der zur Konzeptualisierung und Operationalisierung von Bildungsstandards betrieben werden muss, sie zeigen aber auch den **Gewinn, den die Bildungssteuerung und Unterrichtsforschung aus dieser Arbeit ziehen können**. Die populationsbezogenen Analysen der Items in der Validierungssstudie zeigten etwa markante Unterschiede in den Lesekompetenzen in **Abhängigkeit vom Schultyp auf der Sekundarstufe**, d.h. Klassen mit Grundansprüchen vs. Klassen mit erweiterten und hohen Ansprüchen (HarmoS, 2008: 43). Der Mittelwert beim Leseverstehen in Französisch (Deutschschweiz) lag in den 9. Klassen mit Grundansprüchen tiefer als der Mittelwert in den 6. Klassen der Primarstufe (454 Punkte gegenüber 480 Punkten). Auch in Englisch war die Lesekompetenz in den 6. Klassen der Primarstufe im Durchschnitt besser ausgeprägt als im untersten Niveau der 9. Klassen auf der Sekundarstufe. Ein beträchtlicher Prozentsatz an Lernenden dieses Schultyps scheinen also gewisse Grundkompetenzen, die für eine langfristig erfolgreiche Bildungs- und Berufskarriere unerlässlich sind, nicht ausreichendem Umfang zu erwerben. Genau solche Informationen sind es, welche die aufwändigen Entwicklungsarbeiten zu Bildungsstandards lohnend machen. Die daraus resultierenden Messungen liefern Informationen zur Frage, ob und in welchem Umfang das Bildungssystem seinen Grundauftrag erfüllen kann, nämlich „allen Angehörigen der nachwachsenden Generation – und zwar ausnahmslos – jene Basisqualifikationen zu vermitteln […], die Voraussetzungen für die Teilhabe an gesellschaftlicher Kommunikation und selbstständiges Weiterlernen sind" (Baumert & Kunter, 2006: 475). Bei der Frage, mit welchen Lernformen dies konkret geschehen soll und wie Lehrkräfte auf diese Aufgaben konkret vorbereitet werden können, sind dann wieder Fachdidaktik und Schulforschung gefragt (Keller, 2013a). Dabei muss betont werden, dass auch fachdidaktische Entwicklungs- und Fortbildungsarbeit ohne diese Art von empirischen Hintergrunddaten „im Blindflug" stattfinden.

In jüngster Zeit wird diese Arbeit im Rahmen der Studie **Überprüfung von Grundkompetenzen (ÜGK) fortgesetzt**. Dabei überprüfen die Kantone mit zwei Erhebungen 2016 und 2017 (Ende Primarstufe), wie gut zufällig ausgewählte Gruppen von Schülerinnen und Schüler in der Schweiz einen Ausschnitt der Nationalen Bildungsziele in der Schweiz in der Schulsprache sowie der ersten Fremdsprache erreichen. Die Schülerinnen und Schüler werden auf Tablets getestet, die von externen Testadministrierenden mit in die Schulen gebracht werden. Auch hier werden wiederum – mit Verweis auf die hohen Kosten – nur Lese- und Hörverstehen getestet, während produktive Kompetenzen nicht erfasst werden (vgl. http://uegk-schweiz.ch/).

Wenn sprachproduktive Leistungen wie Schreiben oder Sprechen in den Blick kommen, sind Testformate wie *multiple choice* und damit auch *discrete item tests* –weitgehend ungeeignet. Die standardisierte Beurteilung produktiver Kompetenzen in einem *performance assessment*

Die Arbeiten des Konsortiums zeigen den Aufwand, der zur Konzeptualisierung und Operationalisierung von Bildungsstandards betrieben werden muss, sie zeigen aber auch den Gewinn, den die Bildungssteuerung und Unterrichtsforschung aus dieser Arbeit ziehen kann.

erfordert einen grossen Kodieraufwand, welcher sich daraus ergibt, dass Beobachtungs- bzw. Beurteilungsrichtlinien für Rater erarbeitet und Vorkehrungen getroffen werden müssen, die eine objektive Bewertung der Sprachkompetenzen erlauben (Keller, 2016). Auch ist zunächst eine konzeptuelle Klärung nötig, was überhaupt die Diskurstypen sind, welche beim Assessment erfasst werden sollen. Im Falle der nationalen Bildungsziele beim Schreiben in der Schweiz waren dies die Textfunktionen (a) informieren/beschreiben; (b) auffordern/veranlassen; (c) erzählen/berichten; (d) seine Meinung äussern/argumentieren; sowie (e) Unterhalten von Beziehungen. Sodann gilt es Aufträge zu formulieren, welche eine kommunikative Situation für die Sprachproduktion vorgeben. Bei HarmoS (2007) findet sich das folgende Beispiel: Die Lernenden sollen im Französischunterricht einen Vortrag über Schokolade halten und deshalb bei einer bekannten Firma um Informationen dazu bitten. Dazu sollen sie schreiben, wer sie sind, warum sie schreiben und was sie für ihren Vortrag über Schokolade genau wissen möchten. Diese Beschreibung des eigenen Vorhabens soll in das Genre des Geschäftsbriefs mit korrekter Anrede und Abschluss eingebettet werden.

Durch die Vorgabe solcher Leitpunkte in der Aufgabenstellung entsteht ein sprachpragmatischer Erwartungsraum, für die Texte der Lernenden. Als erfüllt angesehen kann die Aufgabe dann, wenn die entsprechenden Diskurstypen (bei diesem Beispiel (a) und (b)) adressatengerecht umgesetzt sind. Der Aufwand, welcher bei *Discrete Item*-Verfahren in die Validierung der einzelnen Testitems gesteckt wird, fliesst bei einem *Performance Assessment* in die Entwicklung von Beurteilungsrastern, Kodierverfahren und sowie Systemen zur Kontrolle der gleichbleibenden Qualität der Ratings. Prinzipiell kann man dabei holistische oder analytische Raster einsetzen. Beim holistischen Rating besteht der Auftrag an die Rater darin, den Text auf einem einzigen, globalen Kompetenzniveau einzuordnen (z. B. 0 - 5). Dabei können mittels einer (Mehrfacetten-)Rasch-Analyse alle Schülerinnen und Schüler auf einer einzigen, „fairen", Skala abgebildet werden, die der unterschiedlichen Aufgabenschwierigkeit und der Strenge der einzelnen Beurteiler Rechnung trägt.

Für ein analytisches Rating werden die einzelnen Beurteilungskriterien inhaltlich aufgeschlüsselt und konkretisiert, wobei die Rater also einzelne Qualitätsaspekte von Texten separiert erfassen. Damit werden differenziertere Einsichten ermöglicht sowie eine Datenbasis geschaffen, um in vertiefenden Untersuchungen in den Kompetenzprofilen der Lernenden charakteristische Muster zu identifizieren. Die einzelnen Items der analytischen Beurteilungsinstrumente gruppieren sich bei HarmoS (2007) um die folgenden Gesichtspunkte:

1. kommunikative Wirkung (global) und Relevanz des Textes
2. Beachtung und Elaboriertheit der Inhaltspunkte
3. pragmatische Aspekte auf der Textebene:
   - Textlänge
   - Textstrukturierung
   - Themenentwicklung
4. lexikalische Aspekte
   - lexikalisches Spektrum
   - lexikalische Korrektheit
   - linguistische Aspekte der Textstrukturierung (Kohäsion)
5. grammatische Aspekte
   - Verbalbereich
   - Satzbereich (HarmoS 2007: 47)

Bei jeder neuen Aufgabe oder Population von Lernenden, die beurteilt werden sollen, braucht es eine Phase des Ratertrainings. Die Hauptziele davon sind, die Rater zu einem Skalenverständnis zu führen, das der Intention der Verfasser entspricht, und die sichere Anwendung der Kriterien auf unterschiedliche Schülertexte vorzubereiten. Zur Unterstützung solcher Prozesse werden in jüngster Zeit auch computer-basierte Systeme von künstlicher Intelligenz eingesetzt, welche komplementär zu menschlichen Beurteilern verwendet werden und Urteilsfehler wie Reihenfolgeeffekte, Halo- oder Verzerrungseffekte oder Probleme der variierenden Raterstrenge mildern können (Attali, Bridgeman & Trapani, 2010; Keller, 2016). Solche Verfahren erhöhen aber eher die Qualität der *performance assessments*, als dass sie die Kosten senken – sie werden deshalb erst vereinzelt angewandt.

3. Fazit: How much is enough?

Jenseits dieser Prozesse von Aufgaben- und Testentwicklung ist auch die Frage relevant, wann ein Standard als „erfüllt" angesehen werden kann. In Sprache gegossene, mehr oder weniger präzisierte Leistungserwartungen, wie sie in Bildungsstandards festgehalten sind, lassen offen, welche konkreten Leistungen bei der Bewältigung von Aufgaben gezeigt werden müssen. Als Folge sind die häufig geführten Diskussionen über Mindest-, Regel- und Maximalstandards obsolet, so lange keine empirischen Befunde zu tatsächlichen Leistungen von Schülerinnen und Schülern in standardisierten Tests vorliegen (Köller, 2008). Erst die Bereitstellung einer Leistungsskala und die Definition von Kompetenzstufen, wie sie beispielsweise durch das HarmoS Konsortium (2008) vorgenommen wurden, erlauben die finale Festlegung von Cut Scores, bei denen Standards – egal ob Mindest-, Regel- oder Maximalstandards – als erreicht angesehen werden. In diesem Zusammenhang ist es wichtig festzuhalten, dass die definierten Niveaustufen im Gegensatz zu den Kompetenzmodellen, in denen die Dimensionen präzisiert werden, in der Regel nicht wirklich theoriebasiert sind. Sie stützen sich vielmehr auf post hoc Analysen der gewonnenen Daten und sind an einen Standard-Setting-Prozess gekoppelt (vgl. Hambleton, Jaeger, Plake & Mills, 2000), bei dem im Zusammenspiel psychomet-

rischer Analysen, fachdidaktischer Erwägungen und politischer Verträglichkeit gut interpretierbare Bereiche auf dem Kompetenzkontinuum definiert werden. Es handelt sich dabei um konsensuelle Festlegungen, welche die Interpretation von Testwerten erleichtern, die aber weit davon entfernt sind, grundlagenwissenschaftlich fundiert zu sein.

Als Fazit kann man festhalten, dass eine wissenschaftlich fundierte Operationalisierung von **zentralen Kernkompetenzen im Rahmen von Bildungsstandards für die empiriebasierte Weiterentwicklung des Fremdsprachenunterrichts unerlässlich** ist. Dazu gibt es auch kreative und ästhetische Kompetenzen, oder Fähigkeiten im Bereich der komplexen Handlungssteuerung, die sich über standardisierte Tests weniger gut messen lassen. Das heisst aber nicht, dass man sie nicht erfassen oder dokumentieren könnte, z.B. im Rahmen von Portfolios (Keller & König, 2017). Gerade in diesem Bereich der Weiterentwicklung der Lehr- und Lernkultur bietet die Kompetenzorientierung noch Chancen, welche bisher nicht ausreichend genutzt werden. Lehrpersonen, Fachdidaktiker und Bildungsforschende sollten diese Herausforderung gemeinsam annehmen.

Als Fazit kann man festhalten, dass eine wissenschaftlich fundierte Operationalisierung von zentralen Kernkompetenzen im Rahmen von Bildungsstandards für die empiriebasierte Weiterentwicklung des Fremdsprachenunterrichts unerlässlich ist.

Referenzen

Alderson, C. (2000). *Assessing Reading.* Cambridge: Cambridge University Press.

Attali, Y., Bridgeman, B. & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning & Assessment*, 10. Online auf http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603/1455, abgerufen am 18. 1. 2018

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9 (2006) 4, 469-520.

EDK - Schweizerische Konferenz der kantonalen Erziehungsdirektoren (1998). *Gesamtsprachenkonzept.* Online auf http://www.edk.ch/dyn/11671.php, abgerufen am 14.1.2018

EDK – Schweizerische Konferenz der kantonalen Erziehungsdirektoren (2011). *Grundkompetenzen für die Fremdsprachen.* Online auf http://www.edk.ch/dyn/12930.php, abgerufen am 14.1.2018.

EDK – Schweizerische Konferenz der kantonalen Erziehungsdirektoren (2016). *Lehrplan 21.* Online auf https://www.lehrplan.ch/, abgerufen am 14.1.2018.

Ender, S., Moser, U., Imlig, F. & Müller, S. (2017). *Bildungsbericht Nordwestschweiz 2017.* Zürich: Institut für Bildungsevaluation.

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

Hambleton, R., Jaeger, R., Plake, B. & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement* 24, 355–366.

HarmoS – Konsortium HarmoS Fremdsprachen (2008). *Fremdsprachen - Wissenschaftlicher Kurzbericht und Kompetenzmodell.* Online auf https://edudoc.ch/static/web/arbeiten/harmos/L2_wissB_25_1_10_d.pdf, abgerufen am 18.1.2018.

Keller, S. & König, F. (Hrsg.) (2017). *Kompetenzorientierter Unterricht mit Portfolio. Handlungskompetenzen fördern, dokumentieren und beurteilen.* Bern: hep Verlag.

Keller, S. (2016). Measuring Writing at Secondary Level (MEWS). Eine binationale Studie. *Babylonia* 3 /2016, 46-48.

Keller, S. (2013a). *Kompetenzorientierter Englischunterricht.* Berlin: Cornelsen Scriptor.

Keller, S. (2013b). Integrative Schreibdidaktik Englisch für die Sekundarstufe. Theorie, Prozessgestaltung, Empirie. *Giessener Beiträge zur Fremdsprachendidaktik*. Tübingen: Narr Francke Attempto Verlag GmbH.

Klieme, E., Avernarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. (2003). *Expertise. Zur Entwicklung nationaler Bildungsstandards.* Bonn: Bundesministerium für Bildung und Forschung.

Köller, O. (2008). Bildungsstandards - Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. *Zeitschrift für Pädagogik* 54/2, 163-173.

OECD (2003). Reading Literacy. In: OECD (Hrsg.), *The PISA 2003 Assessment Framework*, 107-129. Online auf: http://www.pisa.oecd.org/dataoecd/38/52/33707212.pdf, abgerufen am 18.1.2018.

PISA - Deutsches PISA-Konsortium (Hrsg.) (2007). *PISA '06. Die Ergebnisse der dritten internationalen Vergleichsstudie.* Münster: Waxmann.

Rupp, A., Vock, M., Harsch, C. & Köller, O. (2009). *The development, calibration, and validation of standards-based tests for English as a first foreign language.* Münster: Waxmann.

Weigle, S. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weinert, F. (2001). Concept of competence. A conceptual clarification. In: D. Rychen & L. Hersch Salganik (Hrsg.), *Defining and selecting key competencies.* Göttingen: Hogrefe, S. 45-65.