Zeitschrift für Interkulturellen Fremdsprachenunterricht
Didaktik und Methodik im Bereich Deutsch als Fremdsprache
ISSN 1205-6545 Jahrgang 16, Nummer 2 (Oktober 2011)

Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch

Claudia Harsch & Johannes Hartig

Claudia Harsch
University of Warwick
The Centre for Applied Linguistics
Social Sciences Building
University of Warwick
Coventry, CV4 7AL UK
Tel: +44 (0)24 76575912
Fax: +44 (0)24 76524318
Email: C.Harsch@warwick.ac.uk

Johannes Hartig
Deutsches Institut für
Internationale Pädagogische Forschung
Schloßstraße 29
60486 Frankfurt
Tel: +49 (0)69 24708116
Fax: +49 (0)69 24708444
Email: hartig@dipf.de

Abstract. Fremdsprachliche Kompetenzen können qualitativ auf unterschiedlichen Kompetenzniveaus beschreiben werden, um Informationen über die Bedeutung von unterschiedlichen Ausprägung der Kompetenzen zu geben. Kompetenzniveaus werden auch in Schulleistungstests genutzt, wobei hier empirisch-quantitative Methoden zur Niveaubildung eingesetzt werden. Der vorliegende Beitrag verbindet quantitative Methoden mit qualitativen Beschreibungen, um zu validen Niveaueinteilungen zu kommen. Dazu werden drei typische Vorgehensweisen zur Definition von Kompetenzniveaus diskutiert, die *Post-Hoc-Beschreibung willkürlich gebildeter Niveaus*, die *Ableitung von Kompetenzniveaus aus Aufgabenmerkmalen* und die *Modellbasierte Aufgabenkonstruktion und Niveaudefinition*. Letztere Vorgehensweise wird am Beispiel der Testentwicklung zur Evaluation der Bildungsstandards Englisch illustriert. Abschließend werden Grenzen der vorgestellten Methoden und Implikationen für Unterricht und Forschung diskutiert.

Foreign language proficiency can be qualitatively described at different competency levels in order to convey the meaning of the different levels of proficiency. Such proficiency levels are also used in foreign language testing in schools where empirical-quantitative methods are usually employed to establish the different levels. The present paper aims to combine qualitative and quantitative approaches with the aim of deriving valid level definitions. To achieve this aim, we discuss three typical approaches to establish levels of proficiency, i.e., the *post-hoc description of arbitrarily defined levels*, the *definition of levels based on difficulty-determining task characteristics* and *model-based task construction and level definition*. The latter approach will be illustrated with a description of the test development process employed in the project „Evaluation of the Educational Standards for English as a Foreign Language in Germany“. We conclude our contribution with a discussion of limitations of the three approaches and implications for teaching and research.

Schlagwörter: Definition von Kompetenzniveaus, Aufgabenmerkmale, modellbasierte Aufgabenkonstruktion, Fremdsprachenkompetenz

1. Hintergrund

Fremdsprachliche Kompetenzen können auf unterschiedlichen Kompetenzniveaus qualitativ beschreiben werden, um Lernenden wie Lehrenden Informationen über die Beschaffenheit der Kompetenzen, ihre Ausprägungen und die Anforderungen auf den unterschiedlichen Niveaus zur Verfügung zu stellen. Ein Beispiel für solch eine qualitative Beschreibung von Kompetenzniveaus findet sich im *Gemeinsamen Europäischen Referenzrahmen für Sprachen* (GER, Council of Europe 2001), welcher unterschiedliche kommunikative Aktivitäten und sprachliche Kompetenzen auf sechs Niveaus beschreibt. Kompetenzniveaus sind ebenfalls bekannt aus Schulleistungsstudien, wobei im Bereich des Sprachtestens oft empirisch-quantitative Methoden zur Bildung von Kompetenzniveaus eingesetzt werden. Der vorliegende Beitrag stellt dar, wie quantitative Methoden mit qualitativen Beschreibungen verbunden werden können, um zu möglichst validen Niveaueinteilungen zu kommen. Dazu werden drei typische Vorgehensweisen zur Definition und Beschreibung von Kompetenzniveaus diskutiert, die *Post-Hoc-Beschreibung willkürlich gebildeter Niveaus*, die *Ableitung von Kompetenzniveaus aus Aufgabenmerkmalen* und die *Modellbasierte Aufgabenkonstruktion und Niveaudefinition*. Letztere Vorgehensweise wird anschließend am Beispiel der Testentwicklung zur Evaluation der Bildungsstandards Englisch illustriert, wobei 46 Experten Testaufgaben für die Bereiche Leseverstehen, Hörverstehen und Schreiben auf die Niveaus des GER einschätzten. Zur Bestimmung der Niveaugrenzen kam die konsensbasierte Bookmark-Methode zum Einsatz. Abschließend werden Grenzen der vorgestellten Methoden und Implikationen für Unterricht und Forschung diskutiert.

Zur Erfassung von Fremdsprachkompetenzen kommen in verschiedenen Kontexten *standardisierte Tests* zum Einsatz, d. h. Verfahren, bei denen die Inhalte, Fragen und Bearbeitungsbedingungen für alle getesteten Personen identisch oder weitgehend vergleichbar sind. Beispiele sind internationale Schulleistungsstudien wie PISA oder TIMMS, fremdsprachliche Zertifikate wie etwa die Cambridge Tests oder die DELF und DALF Zertifikate, oder Studien wie DESI oder die Evaluation der Bildungsstandards, die im Bereich des nationalen Bildungsmonitoring Verwendung finden.

Aus der Auswertung standardisierter Tests resultieren in der Regel numerische Testwerte, die Auskunft über die individuellen Ausprägungen der getesteten Personen im zu testenden Merkmal, z. B. der Lesekompetenz in einer Fremdsprache, geben sollen. Mögliche einfache Varianten von Testwerten aus standardisierten Tests sind die Menge oder der Anteil gelöster Aufgaben (z. B. „24 Aufgaben richtig beantwortet“ oder „74% der Aufgaben richtig beantwortet“). Derartige *Rohwerte* haben für sich genommen zunächst wenig Aussagekraft, da sie nicht nur von der zu messenden Kompetenz abhängen, sondern auch von der Schwierigkeit der für den Test ausgewählten Aufgaben. Um ein individuelles Testergebnis interpretieren zu können, ist ein *Vergleichsmaßstab* erforderlich, hinsichtlich dessen das Ergebnis eingeordnet werden kann. Mit Bezug auf verschiedene Vergleichsmaßstäbe unterscheidet man in der Diagnostik zwischen einer *normorientierten Testwertinterpretation* und einer *kriteriumsorientierten Testwertinterpretation* (vgl. Goldhammer & Hartig 2007).

Bei einer normorientierten Testwertinterpretation wird das Testergebnis in Bezug zur Verteilung der Testergebnisse einer Bezugsgruppe gesetzt (man spricht hierbei auch von einer „sozialen Bezugsnorm“). Die Interpretation des Testergebnisses bezieht sich dann darauf, ob die individuellen Leistungen im Test z. B. im Vergleich mit anderen Personen (z. B. Schülerinnen und Schülern derselben Jahrgangsstufe) als unter- oder überdurchschnittlich einzuordnen sind. Eine normorientierte Interpretation eines Ergebnisses eines Leseverstehenstests kann sich z. B. darauf beziehen, dass das Testergebnis eines Schülers nur von 25% der Schülerinnen und Schüler derselben Jahrgangsstufe übertroffen wurde. Häufig werden Testergebnisse aus standardisierten Tests, um eine normorientierte Interpretation zu erleichtern, in sogenannte *Normwerte* transformiert, aus denen die Einordnung des Testergebnisses direkt hervorgeht (zu verschiedenen Normwerten s. Goldhammer & Hartig, 2007). Ein Beispiel für Normwerte sind die für die PISA-Studien verwendeten Testwertskalen: Die Messwerte für die in PISA erfassten Kompetenzen werden so transformiert, dass 500 der mittleren Leistung der 15jährigen in den OECD-Staaten entspricht und 100 einer Standardabweichung in dieser Population (OECD 2009).

Für die Diagnostik in pädagogischen Kontexten, so auch oft in der Fremdsprachdiagnostik, ist die Interpretation anhand von sozialen Bezugsnormen oft unzureichend. Es interessiert vielmehr, über welche konkreten Kompetenzen

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

eine getestete Person verfügt, welche fachlichen Anforderungen sie bewältigen kann und welche (noch) nicht. Eine derartige, auf inhaltlich definierte Kriterien bezogene Interpretation bezeichnet man als kriteriumsorientierte Testwertinterpretation. Eine kriteriumsorientierte Interpretation eines Ergebnisses eines Leseverstehenstests kann sich z. B. darauf beziehen, dass ein getesteter Schüler in der Lage ist, gezielt Informationen aus anspruchsvollen Texten zu entnehmen, aber noch nicht imstande ist, Interpretationen hinsichtlich der Intention des Autors vorzunehmen. Bei der kriteriumsorientierten Testwertinterpretation eines individuellen Ergebnisses spielt es keine Rolle, wie viele andere Personen ein genauso gutes, besseres oder schlechteres Testergebnis erzielt haben.

In der Regel ist es nicht möglich, bei einer kriteriumsorientierten Interpretation inhaltlich so fein zu differenzieren, wie dies anhand eines numerischen Testergebnisses technisch möglich wäre, also z. B. anzugeben, was den Unterschied zwischen der Interpretation von 532, 533 oder 534 Punkten auf der PISA-Skala für Lesekompetenz ausmacht.

Um dennoch eine inhaltliche, kriteriumsorientierte Beschreibung von Testergebnissen vorzunehmen, werden häufig sogenannte Kompetenzniveaus oder Kompetenzstufen[1] verwendet. Dies bedeutet, dass die Testwertskala in Abschnitte unterteilt wird und für diese dann eine inhaltliche Beschreibung vorgenommen wird, z. B. beschreibt die „Kompetenzstufe II“ in PISA 2009 den Wertebereich von 408 bis 479 Punkten auf der internationalen Skala für Lesekompetenz (vgl. Naumann, Artelt, Schneider & Stanat 2010). Für Schülerinnen und Schüler mit einem Punktwert in diesem Bereich wird davon ausgegangen, dass man ihre Lesekompetenz angemessen mit den in der Beschreibung für Kompetenzstufe II zusammengefassten Kriterien beschreiben kann (z. B. „… können innerhalb eines Textabschnittes logischen und linguistischen Verknüpfungen folgen, mit dem Ziel, Informationen im Text zu lokalisieren oder zu interpretieren“; Naumann et al. 2010: 28). Am Beispiel der PISA-Skala wird deutlich, dass sich eine norm- und eine kriteriumsorientierte Interpretation nicht ausschließen, sondern Messwerte einer Skala in beiderlei Hinsicht interpretiert werden können.

Bei der Definition von Kompetenzniveaus für ein Testverfahren sind zwei wichtige Fragen zu klären:

(1) Wo werden die Grenzen zwischen verschiedenen Niveaus gesetzt (und damit einhergehend: wie viele Niveaus werden definiert)?

(2) Wie werden die inhaltlichen Beschreibungen für die Kompetenzniveaus vorgenommen?

Zur Definition von Kompetenzniveaus gibt es eine Vielzahl unterschiedlicher Verfahrensweisen, die sich hinsichtlich der Antworten auf diese beiden Fragen unterscheiden. Der vorliegende Beitrag hat zunächst zum Ziel, unterschiedliche Vorgehensweisen zur Definition von Kompetenzniveaus darzustellen. Darauf aufbauend wird das Vorgehen bei den Bildungsstandards in Englisch für die Sekundarstufe I dargestellt, das eines der im Folgenden dargestellten Verfahren genauer illustriert.

Zunächst wird kurz auf die Item-Response-Theorie (IRT) eingegangen, die eine wesentliche methodische Grundlage für die meisten Verfahren zur Definition von Kompetenzniveaus darstellt. Anschließend werden drei verschiedene typische Strategien zur Definition von Kompetenzniveaus betrachtet, nämlich (1) die *Post-Hoc-Beschreibung willkürlich gebildeter Niveaus*, (2) die *Ableitung von Kompetenzniveaus aus Aufgabenmerkmalen* und (3) die *Modellbasierte Aufgabenkonstruktion und Niveaudefinition*. Schließlich erfolgt eine ausführliche Beschreibung des Vorgehens für die Bildungsstandards in Englisch für die Sek. I.

1.1 Die Item-Response-Theorie als Grundlage für eine kriteriumsorientierte Testwertinterpretation

Die Item-Response-Theorie (IRT) (Moosbrugger 2007) stellt eine Sammlung von modernen Methoden zur statistischen Auswertung von Daten aus standardisierten Testverfahren zur Verfügung. Mit diesen Methoden werden eine Reihe technischer Probleme gelöst, die sich bei der Auswertung von Testdaten mit älteren Verfahren im Rahmen der sogenannten klassischen Testtheorie (KTT) ergeben. So können z. B. auch dann für alle getesteten Personen vergleichbare Messwerte erzeugt werden, wenn diese unterschiedliche Teilmengen von Testaufgaben bearbeitet haben. Der für die Definition von Kompetenzniveaus wichtigste Vorteil der IRT ist die Definition einer *gemeinsamen Skala*, auf der sowohl die Messwerte für die individuellen Kompetenzen als auch die Schwierigkeiten der Testaufgaben

---

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

abgebildet werden können. Das Erstellen dieser gemeinsamen Skala durch Anwendung der IRT auf empirische Daten wird auch als *Skalierung* bezeichnet.

Kern der IRT sind statistische Modelle, die den Zusammenhang zwischen der zu messenden Kompetenz und der *Lösungswahrscheinlichkeit* für einzelne Testaufgaben darstellen. Das in unserem Kontext wichtigste IRT-Modell, auf das wir uns im Weiteren beschränken werden, ist das Raschmodell (Rasch 1960). Im Raschmodell werden Annahmen über die Zusammenhänge zwischen den Lösungswahrscheinlichkeiten für Testaufgaben in Abhängigkeit von der Schwierigkeit der Aufgaben und der Kompetenz der getesteten Personen formuliert. In Abbildung 1 ist dieser Zusammenhang exemplarisch für drei prototypische Testaufgaben für Englisch Lesekompetenz aus der DESI-Studie dargestellt. Die Schwierigkeit einer Aufgabe ist im Raschmodell mit Bezug auf diese Lösungswahrscheinlichkeiten definiert: Eine Aufgabe, die sich an einer bestimmten Stelle auf der Raschskala befindet, kann von einer Person, deren Kompetenz sich auf derselben Stelle der Raschskala befindet, mit einer Wahrscheinlichkeit von 50% gelöst werden. Da eine Lösungswahrscheinlichkeit von 50% zu niedrig ist, um Personen eine „hinreichende Sicherheit“ zur Bewältigung bestimmter Aufgaben zu attestieren, werden die aus dem Raschmodell ermittelten Schwierigkeiten häufig auf der Testwertskala so verschoben, dass sie sich auf eine höhere Schwelle als 50% beziehen (z. B. 62% in den PISA-Studien oder 65% in der DESI-Studie; zum Vorgehen vgl. Rauch & Hartig 2007). Für Abbildung 1 sind die Schwierigkeiten der drei Aufgaben so gewählt, dass die Schwierigkeiten typischen Anforderungen der Kompetenzniveaus A, B und C für Englisch Leseverstehen in der DESI-Studie entsprechen. Auf Niveau A können konkrete Einzelinformationen aus sprachlich einfachen Texten entnommen werden, auf Niveau B kann eine begrenzte Menge von Informationen aus einfachen Texten verknüpft und interpretiert werden und auf Niveau C kann komplexe Information aus sprachlich anspruchsvollen Texten inferiert und verknüpft werden (vgl. Nold; Rossa & Chatzivassiliadou 2008).

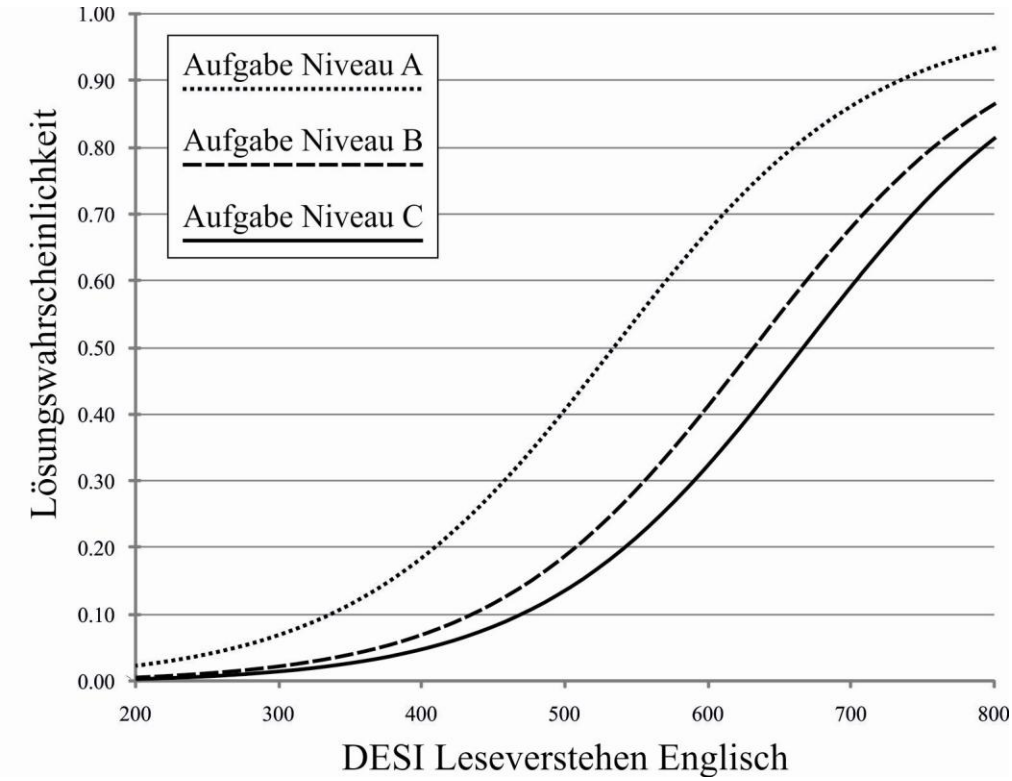

Abbildung 1: Aus dem Raschmodell vorhergesagte Lösungswahrscheinlichkeiten für drei prototypische Aufgaben für Niveau A, B und C in Englisch Leseverstehen in Abhängigkeit von der Leseverstehens-Kompetenz der Schülerinnen und Schüler.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

Wie in Abbildung 1 veranschaulicht wird, lassen sich aus dem Raschmodell die Wahrscheinlichkeiten vorhersagen, mit denen Personen mit einer bestimmten Kompetenz bestimmte Testaufgaben lösen. So werden z. B. Personen mit 600 Punkten auf der Skala in Abbildung 1 das erste Item mit recht hoher Wahrscheinlichkeit (68%) lösen, das zweite nur noch mit einer Wahrscheinlichkeit von 42% und das dritte wahrscheinlich nicht (nur mit einer Wahrscheinlichkeit von 33%). Diese durch eine IRT-basierte Auswertung mögliche Übertragung von Testwerten in Lösungswahrscheinlichkeiten stellt die Ausgangsbasis für eine kriteriumsorientierte Testwertinterpretation dar. Durch eine gemeinsame Betrachtung der Anforderungen der Testaufgaben mit den erwarteten Lösungswahrscheinlichkeiten der Aufgaben wird es nämlich möglich, für Personen mit bestimmten Testwerten Aussagen darüber zu machen, welche Anforderungen sie schon mit hinreichender Wahrscheinlichkeit bewältigen können und welche (noch) nicht.

2. Strategien zur Konstruktion von Kompetenzniveaus

In den folgenden Abschnitten werden die drei oben genannten Strategien zur Definition von Kompetenzniveaus (Post-Hoc-Beschreibung, Bezug auf Aufgabenmerkmale, und Modellbasierte Aufgabenkonstruktion und Niveaudefinition) beschrieben. Wir haben diese drei Strategien ausgewählt, da sie typische Vorgehensweisen darstellen, anhand derer die Vielfalt möglicher Vorgehensweisen veranschaulicht werden soll. In unserem Beitrag können wir keine erschöpfende Darstellung der Vielfalt der in verschiedenen Studien verwendeten Verfahren bieten. Die Grenzen zwischen den hier dargestellten Strategien sind fließend, sie unterscheiden sich jedoch hinsichtlich des Vorwissens, das in die Konstruktion der Testverfahren und in die Definition der Kompetenzniveaus einfließt. Dementsprechend sind sie für unterschiedliche Kompetenzbereiche je nach Forschungsstand unterschiedlich gut geeignet.

2.1 Post-Hoc-Beschreibung willkürlich gebildeter Niveaus

Ein mögliches Vorgehen zur Definition von Kompetenzniveaus besteht darin, zunächst nach willkürlichen oder externen Kriterien Schwellenwerte zwischen Niveaus (und damit auch die Anzahl der Niveaus) festzulegen, z. B. in gleichen Abständen auf der Testwertskala oder orientiert an den durchschnittlichen Leistungen von Schülern unterschiedlicher Jahrgangsstufen. Nach der Setzung der Schwellen werden dann Testaufgaben gesucht, die aufgrund ihrer Schwierigkeiten geeignet sind, die Unterschiede zwischen den Niveaus zu charakterisieren, da sie z. B. auf einem bestimmten Niveau mit hinreichender Wahrscheinlichkeit bewältigt werden können und auf dem darunter liegenden Niveau noch nicht. Die Inhalte der so ausgewählten Aufgaben werden dann post hoc analysiert und aus der Analyse werden Beschreibungen der Kompetenzniveaus abgeleitet. Für eine systematische Beschreibung des hier skizzierten Vorgehens siehe Beaton & Allen (1992). Die Post-Hoc-Beschreibung von Niveaus ist insbesondere dann ein angemessenes Vorgehen, wenn noch wenig gesichertes Vorwissen über die Anforderungen der Testaufgaben verfügbar ist und noch kein theoretisches Modell über die Niveaustruktur der zu messenden Kompetenz vorliegt. Ein Nachteil dieses Vorgehens ist der fehlende inhaltliche Bezug auf das zu messende Konstrukt bei der Schwellensetzung, auch Anzahl und Breite der Niveaus sind willkürlich. Zudem entsteht bei der post hoc vorgenommenen Beschreibung der Aufgabeninhalte ein gewisser Spielraum für willkürliche Interpretationen.

2.2 Ableitung von Kompetenzniveaus aus Aufgabenmerkmalen

Eine empirisch fundierte Definition von Kompetenzniveaus kann durch die Nutzung von Informationen über die Anforderungen der Testaufgaben vorgenommen werden. Voraussetzung ist, dass alle Testaufgaben gleichermaßen hinsichtlich relevanter Anforderungen, sogenannter *schwierigkeitsbestimmende Merkmale*, beschrieben werden. Schwierigkeitsbestimmende Merkmale für Testaufgaben zur Lesekompetenz in einer Fremdsprache können z. B. die sprachlichen Anforderungen der Lesetexte (Vokabular, Grammatik) sein oder die Komplexität der kognitiven Prozesse, die die Fragen fordern (z. B. einfache Informationen entnehmen vs. Interpretieren). Liegen Beschreibungen der Testaufgaben hinsichtlich derartiger Merkmale vor, kann folgendes Vorgehen zur Definition von Kompetenzniveaus gewählt werden (vgl. Hartig 2007):

1) Die Aufgabenschwierigkeiten werden auf Basis empirischer Daten mit dem Raschmodell geschätzt.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

2) Die geschätzten Schwierigkeiten werden in einem Regressionsmodell mit den Aufgabenmerkmalen vorhergesagt.

3) Aus den Ergebnissen der Regressionsanalyse werden *erwartete Aufgabenschwierigkeiten* für Aufgaben mit spezifischen Kombinationen von Anforderungen ermittelt. Diese erwarteten Aufgabenschwierigkeiten können als Punkte auf der Testwertskala interpretiert werden, an denen getestete Personen Aufgaben mit spezifischen Anforderungen mit einer hinreichenden Sicherheit lösen können.

4) Die erwarteten Aufgabenschwierigkeiten für ausgewählte Anforderungskombinationen werden als Schwellen zur Unterscheidung von Kompetenzniveaus verwendet. Die inhaltliche Beschreibung der Niveaus erfolgt durch Bezug auf die jeweiligen Aufgabenanforderungen.

Das hier skizzierte Verfahren (für eine ausführliche Beschreibung s. Hartig 2007) reduziert die Willkür der im vorangehenden Abschnitt beschriebenen Methode deutlich. Es erlaubt es, ein Niveaumodell zu konstruieren, das durch die Effekte der Aufgabenmerkmale empirisch fundiert ist. Notwendige Voraussetzung für eine erfolgreiche Anwendung des Vorgehens ist allerdings, dass die Aufgabenmerkmale tatsächlich zur Vorhersage der empirischen Schwierigkeiten geeignet sind. Zudem hängt die Güte des damit konstruierten Kompetenzmodells wesentlich davon ab, dass bei der Beschreibung der Testaufgaben keine wichtigen Anforderungen außer Acht gelassen wurden.

2.3 Modellbasierte Aufgabenkonstruktion und Niveaudefinition

Wenn ein hinreichend ausformuliertes *Niveaumodell* für ein zu messendes Konstrukt existiert, kann bereits die Aufgabenkonstruktion an diesem Modell ausgerichtet werden. Typischerweise werden in einem Kompetenzniveaumodell allgemeine Anforderungen beschrieben, die Personen auf unterschiedlichen Niveaus bewältigen können. Diese Beschreibungen müssen in konkrete Spezifikationen zur Entwicklung von Testaufgaben übersetzt werden, in denen formuliert wird, mit welchen Arten von Testaufgaben die für verschiedene Kompetenzniveaus typischen Anforderungen getestet werden können. Den Aufgaben, die bezogen auf derartige modellbasierte Spezifikationen zur Testentwicklung konstruiert wurden, können *a priori angenommene Schwierigkeiten* zugeordnet werden – Aufgaben, die die Spezifikationen eines niedrigeren Niveaus erfüllen, sollten z. B. leichter sein als solche, die Spezifikationen eines höheren Niveaus erfüllen.

Sind Aufgaben mit Bezug auf ein Modell konstruiert worden, werden ihre empirischen Schwierigkeiten durch eine Skalierung auf Basis von Testdaten ermittelt – diese empirischen Schwierigkeiten werden in der Regel nicht vollkommen mit den a priori angenommenen übereinstimmen (für eine Diskussion möglicher Gründe siehe auch Abschnitt 0). Die aus der Skalierung resultierende Testwertskala muss dann in die Kompetenzniveaus des Modells unterteilt werden, um Personen den Niveaus zuordnen zu können. Dazu werden die Schwellen zwischen den Niveaus auf der Skala bestimmt. Dies erfolgt in der Regel durch sogenannte *Standard-Setting-Verfahren*. Beim Standard Setting werden Expertenurteile benutzt, um die IRT-basierten Testwertskalen in Kompetenzniveaus zu unterteilt. Dabei gibt es eine Vielzahl an möglichen Verfahren (für eine Übersicht vgl. Cizek, Bunch & Koons 2004; Zieky & Perie 2006). Grundsätzlich können zwei Gruppen von Verfahren unterschieden werden: *Testorientierte Verfahren (test-centred approaches)* nutzen die skalierten Aufgabenschwierigkeiten als Basis zur Niveaueinteilung. *Probandenorientierte Verfahren (examinee-centred approaches)* beziehen sich auf die Kompetenzen getesteter Personen, um die IRT-Skala in Niveaus einzuteilen. Bei diesen Verfahren werden z. B. Lehrkräfte benötigt, die die getesteten Schüler gut genug kennen. Ist die Aufgabenkonstruktion an einem Kompetenzmodell orientiert erfolgt, bieten sich testorientierte Standard-Setting-Verfahren an, bei denen die Experten die Aufgaben bezüglich der Niveaus des Modells einschätzen. Aus diesen Experteneinschätzungen lassen sich die Grenzen zwischen Kompetenzniveaus bestimmen. Das modellbasierte Vorgehen wird im Folgenden anhand der Bildungsstandards für die erste Fremdsprache Englisch (Sek. 1) dargestellt.

3. Kompetenzniveaus der Bildungsstandards

Im Folgenden soll ein Verfahren illustriert werden, das den oben beschriebenen Weg einer an einem existenten Kompetenzmodell orientierten Testentwicklung und Niveaudefinition geht. Die Kriterien zur Interpretation werden

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

dabei ebenso vom Referenzkompetenzmodell abgeleitet wie die Aufgaben-Charakteristika zur Testentwicklung. Wir beschreiben das Vorgehen im Projekt der Evaluation der Bildungsstandards der KMK (KMK 2003; KMK 2004) in der Fremdsprache Englisch für die Testbereiche des Schreibens, und des Lese- und Hörverstehens, das in den Jahren 2006 bis 2010 am Institut zur Qualitätsentwicklung im Bildungswesen, Humboldt-Universität zu Berlin durchgeführt wurde. Als Referenz wurde das Kompetenzmodell des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER, Council of Europe 2001) genutzt, das auch den Bildungsstandards zugrunde liegt. Dieses Modell beschreibt fremdsprachliche Kompetenzen in verschiedenen, hierarchisch gegliederten Bereichen (mittels so genannter Kompetenzskalen) auf sechs ansteigenden Kompetenzniveaus (A1 bis C2). Für die Bildungsstandards wurde das Niveau A2 für den Hauptschulabschluss und die Niveaus B1/B2 für den Mittleren Schulabschluss angesetzt. Die Deskriptoren der Niveaubeschreibungen im GER und in den Bildungsstandards zielen auf eine positive Beschreibung des Könnens der Lernenden ab.

3.1 Vorgehen bei der Niveaudefinition

Das Kompetenzmodell des GER wurde genutzt als Ausgangspunkt der Testentwicklung und als Zielbeschreibung der Kompetenzniveaus. Nun ist der GER kein Instrument der Testentwicklung und hat auf diesem Gebiet erkennbare Defizite (vgl. Harsch 2007; Weir 2005). Beispielsweise sind im GER keine Testaufgaben oder deren relevante Merkmale beschrieben; die GER-Deskriptoren sind nicht in allen Bereichen kohärent verfasst; ihre Funktion ist die einer generellen, benutzerorientierten Kompetenzbeschreibung, nicht die einer Konstruktions- oder Bewertungsskala (zur Skalentypologie vgl. Alderson 1991). Deshalb müssen die GER-Deskriptoren zur Testentwicklung zunächst in ein Testkonstrukt und in Testspezifikationen übersetzt werden, die für den gegebenen Kontext relevant sind und den Testentwicklern als Leitfaden dienen können. Ehe dann die Testwerte mithilfe der GER-Deskriptoren und Niveaus interpretiert werden können, müssen die Testaufgaben wiederum in ihren Anforderungen und Schwierigkeiten in Bezug auf die GER-Niveaus und Kriterien eingeschätzt werden und empirisch mittels Standard-Setting angebunden werden. Das *Manual for Relating Examinations to the Common European Framework* (Council of Europe 2009) gibt einen hilfreichen Überblick über eine Reihe solcher Verfahren: es wurde auch im hier beschriebenen Projekt zu Grunde gelegt. Im Folgenden beschreiben wir die relevanten Schritte von der Testentwicklung bis hin zur Testwertinterpretation.

3.1.1 Testkonstrukt und Spezifikationen:

Zunächst wurden die Bildungsstandards und die Skalen des GER auf ihre Relevanz für die Testbereiche des Schreibens, und des Lese- und Hörverstehens hin analysiert. Auf dieser Basis wurden für jedes der im Projekt anvisierten Kompetenzniveaus (GER-Niveaus A1 bis C1, um auch Kompetenzbereiche unterhalb und oberhalb der Bildungsstandards abzudecken) Testkonstrukte entwickelt, d.h. eine kriterienorientierte Definition dessen, was beispielsweise Schreiben auf einem bestimmten Niveau kennzeichnet. Diese Konstrukte wurden ergänzt um detaillierte Spezifikationen hinsichtlich der Aufgabenanforderungen und Charakteristika, wie etwa der geforderten Textlänge in den Schreibaufgaben, der Sprechgeschwindigkeit des Inputs beim Hörverstehen, oder des Sprachniveaus der Texte, die den Leseverstehensaufgaben zugrunde gelegt werden. Im nächsten Schritt wurden auf Basis dieser Konstrukte und Spezifikationen durch herzu geschulte Englisch-Lehrkräfte Testaufgaben entwickelt. Alle Testaufgaben wurden in ihren Anforderungen charakterisiert und von den Aufgabenentwicklern auf das GER-Niveau hin eingeschätzt, das ein Lernender mindestens erreicht haben muss, um diese Aufgabe mit hinreichender Wahrscheinlichkeit zu lösen. Bei dieser a-priori Einschätzung wurden existente Instrumente aus dem Umfeld des GER genutzt, wie etwa das Kriterienraster des Dutch Grid zur Charakterisierung rezeptiver Testaufgaben (Alderson et al. 2006), oder die ALTE-Checklisten zur Charakterisierung von Schreibaufgaben (vgl. ALTE Members 2007). Für eine detaillierte Darstellung des Testentwicklungsprozesses vgl. Rupp et al. (2008).

3.1.2 Testanalyse

Alle Aufgaben wurden in kleinem Rahmen empirisch vorerprobt, gegebenenfalls überarbeitet und dann an einer größeren Stichprobe pilotiert. Die Testdaten wurden mittels des Raschmodells (s. o.), skaliert und analysiert. Die Analysen orientierten sich an international gültigen Teststandards (Details in Rupp & Porsch 2010). Auf Grund der

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

Pilotierungsergebnisse wurden gut funktionierende Aufgaben in die Datenbank eingespeist, während Aufgaben, die sich nicht erwartungsgemäß verhielten, entweder revidiert oder ausgeschlossen wurden. Für die funktionierenden Aufgaben lagen aus dem Raschmodell geschätzte empirische Schwierigkeitswerte vor. Diese konnten nun mit den a-priori eingeschätzten Anforderungen und GER-Niveaueinschätzungen seitens der Aufgabenentwickler verglichen werden: Bei allen Kompetenzbereichen zeigten sich lineare Zusammenhänge, wobei bei den rezeptiven Kompetenzen etwa 40% der empirischen Aufgabenschwierigkeiten erklärt werden konnten (vgl. Leucht, Harsch, & Köller, in Revision). Die a-priori eingeschätzten Schreibaufgaben wurden in ihrer Schwierigkeitsrangfolge durch die empirischen Analysen bestätigt (vgl. Harsch & Rupp 2011). Während diese Ergebnisse ein vielversprechender erster Schritt sind in Richtung auf eine Anbindung der Testaufgaben an die GER-Niveaus, müssen sie dennoch durch die erwähnten formalen Standard-Setting Verfahren untermauert und validiert werden. Dies ist eine der Voraussetzungen, um Schülerleistungen kriterienorientiert auf GER-Niveaus interpretieren zu können.

3.1.3 Standard Setting

Beim Standard Setting werden, wie oben erwähnt, die Testwertskalen in Kompetenzniveaus unterteilt. Da die Testaufgaben für die Evaluation der Bildungsstandards bereits an einem Niveaumodell orientiert entwickelt wurden und die Voraussetzungen für ein personenzentriertes Standard Setting nicht gegeben waren, wurden testorientierte Verfahren gewählt. Für eine detaillierte Beschreibung und Berichtlegung der Standard-Setting Prozeduren in diesem Projekt vgl. Harsch & Tiffin-Richards (2010).

Da der Aufgabenpool zur Evaluation der Bildungsstandards zu viele Aufgaben umfasste, als dass alle in das Standard Setting einfließen könnten, musste vorab eine repräsentative Auswahl getroffen werden. Hierbei wurden relevante Auswahlkriterien genutzt, wie etwa die a-priori Einschätzung durch die Aufgabenentwickler, Testformate, Inhalte, Testkonstrukte und die empirischen Schwierigkeiten, um den gesamten Bereich der erfassten Skala abzudecken.

In der Regel werden die Niveauschwellen von einem Team von Experten in konsensuellen Verfahren festgelegt. Dabei ist es wichtig, dass dieses Team auch Interessenvertreter aus relevanten Bereichen repräsentiert. Im vorliegenden Projekt wurden Experten aus den folgenden Bereichen eingeladen: Bildungsministerien der Länder, Qualitätsinstitute der Länder, praktizierende Lehrkräfte, Akademiker aus den Bereichen der Didaktik, des Sprachtestens und der Psychometrie. Insgesamt 46 Experten aus Deutschland, Österreich, der Schweiz und Norwegen nahmen am Standard Setting teil.

Aufgabe der Experten war es, die Grenzen zwischen benachbarten Kompetenzniveaus durch die so genannte Bookmark Methode (vgl. etwa Mitzel, Lewis, Patz & Green 2001; Tiffin-Richards 2010) festzulegen. Dazu wurden den Experten die Testaufgaben angeordnet nach ihren empirischen Schwierigkeiten vorgelegt: die Experten mussten entscheiden, welche Testaufgaben die Grenze zwischen zwei benachbarten Niveaus darstellen. Dabei wurden die Deskriptoren des GER als Entscheidungsbasis herangezogen: Die Experten mussten die ihnen vorgelegten Testaufgaben mit den Anforderungen und Könnensbeschreibungen der GER-Niveaus abgleichen, um die Niveaugrenzen kriteriengestützt anhand des GER-Kompetenzmodells zu bestimmen.

Das Ergebnis dieses Verfahrens ist eine empirisch gewonnene Kompetenzskala für jeden der drei getesteten Kompetenzbereiche. Diese Kompetenzskalen sind unterteilt in die GER-Kompetenzniveaus A1 bis C1 und sie erlauben es, Aufgabenschwierigkeiten und Schülerkompetenzen auf derselben Skala abzubilden.

3.1.4 Berichten der Schülerkompetenzen und Interpretation der Testwerte

Solche Kompetenzskalen ermöglichen eine kriterienorientierte Interpretation der von den Schülern erzielten Testwerte. Die Niveaus einer Kompetenzskala sind beschrieben durch die Kriterien der GER-Deskriptoren, so dass Testwerte inhaltlich interpretiert werden können, und einem bestimmten Testwert eine kriterienorientierte Könnensbeschreibung zugeordnet werden kann.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

Der Vorteil dieses Vorgehens besteht darin, dass Kompetenzniveaus nicht mehr oder minder „willkürlich“ bestimmt werden, sondern dass diese Niveaus und ihre inhaltlich-kriterielle Beschreibung a priori existieren, zur Testerstellung genutzt werden, als Basis für die Niveaueinteilung der Kompetenzskalen dienen und letztlich wieder zur Testwertinterpretation herangezogen werden. Dieses Verfahren ermöglicht Transparenz von der Beschreibung der Testaufgaben bis hin zur Interpretation der erzielten Testwerte, ausgerichtet an ein und demselben Kompetenzmodell.

3.2 Grenzen dieses Vorgehens

Neben den genannten Vorteilen sollten allerdings gewisse Grenzen solch einer Kompetenzniveaubildung nicht ignoriert werden. Im Folgenden thematisieren wir drei wichtige Aspekte, die als mögliche Ursachen für Grenzen und Unschärfen dieses Verfahrens identifiziert werden können.

3.2.1 „Übersetzungsfehler“ bei der Testentwicklung

Wie oben erläutert, müssen die GER-Deskriptoren, die das Können der Lernenden in genereller Form beschreiben, zunächst in Testspezifikationen „übersetzt“ werden. Im nächsten Schritt werden die Spezifikationen operationalisiert, in konkrete Testaufgaben „übersetzt“. Diese Aufgaben werden dann von den Entwicklern in ihren Merkmalen charakterisiert, welche wiederum auf die Testspezifikationen rückbezogen sind. Auf dieser Charakterisierung basiert die a-priori Schwierigkeitseinschätzung der Testaufgaben in Bezug auf das GER-Niveau, das ein Lerner erreicht haben muss, um die Aufgabe zu lösen. Bei all diesen Schritten nun kann es zu „Übersetzungsfehlern“ kommen, da bei jedem Schritt subjektive Einschätzungen und Interpretationen nötig sind. Diese „Übersetzungsfehler“ sind eine der Ursachen, warum die o.g. a-priori GER-Einstufungen der Entwickler nur etwa 40% der empirischen Schwierigkeitsunterschiede erklären können. Solche Übersetzungsfehler können nicht gänzlich vermieden werden und liegen in der Natur des Testentwicklungsprozesses. Sie können jedoch aufgefangen werden durch die empirischen Analysen der Tests und durch die genannten formalen Standard-Setting Verfahren.

3.2.2 Unschärfen beim Standard-Setting Verfahren

Auch bei den Standard-Setting Verfahren gibt es jedoch Unsicherheiten und Grenzen, da die Experteneinschätzungen wiederum subjektiver Natur sind, und es hier ebenfalls zu „Übersetzungsfehlern“ beim Abgleich der GER-Deskriptoren mit den Anforderungen der Testaufgaben kommen kann: Experten könnten beispielsweise die GER-Deskriptoren unterschiedlich interpretieren, oder die Testaufgaben in ihren Anforderungen über- oder unterschätzen. Dazu tritt das Problem, dass die Experteneinschätzung sich nicht notwendigerweise mit den empirisch ermittelten Schwierigkeiten der Testaufgaben decken muss. Dies könnte seine Ursache in einer Über- oder Unterschätzung der Schwierigkeitsanforderungen oder des Könnens der Schülerpopulation haben. Eine weitere Ursache solcher Diskrepanzen könnte in einer konzeptionellen Vermischung des kontextunabhängigen GER und der deutschen Situation liegen: Zum Beispiel könnte eine Expertenmeinung sein, dass eine bestimmte Aufgabe von einem Realschüler der Klasse 10 nicht gelöst werden kann, weshalb die Aufgabe nicht dem Kompetenzniveau B1 entsprechen kann. Dies ist jedoch eine Umkehrung des geforderten Vorgehens, bei dem von den Anforderungen der Testaufgabe ausgegangen werden müsste: Decken sich die Anforderungen einer Aufgabe mit den GER-Deskriptoren eines gegebenen Niveaus, so sollte ein Lerner dieses Niveaus die Aufgabe lösen können, unabhängig davon, ob ein Realschüler in der deutschen Schule das tatsächlich kann. Auf diesen Aspekt wird im Folgenden noch näher eingegangen.

Diese Unschärfen in der Experteneinschätzung können durch verschiedene Maßnahmen kontrolliert und aufgefangen werden, wie etwa durch die Ermittlung der Inter-Rater Reliabilitäten, oder durch den Einsatz von konsensbasierten Verfahren, bei der sich eine Expertengruppe auf Niveaugrenzen einigen muss, oder durch IRT-basierte Verfahren, bei denen individuelle Unterschiede der Experten mithilfe eines probabilistischen Modells korrigiert werden können.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

3.2.3 Unerwartete Effekte von Aufgabenformaten

Neben den beiden gerade diskutierten Quellen der Unsicherheit kann eine Diskrepanz zwischen antizipierten und empirischen Aufgabenschwierigkeiten ihre Ursache auch in spezifischen Aufgabenanforderungen haben, die im zugrunde gelegten Niveaumodell nicht enthalten sind. Es kann vorkommen, dass Lernende, die Stufe B1 des GER eigentlich erreicht haben, Aufgaben, die auf diesem Niveau verortet wurden, dennoch nicht oder nur selten lösen. Eine mögliche Erklärung für dieses Verhalten kann darin liegen, dass bestimmte kognitive Operationen, die für das Beantworten von Aufgaben mit spezifischen Antwortformaten erforderlich sind (wie etwa das Interpretieren vorstrukturierter Informationen beim Ergänzen von Tabellen (*table filling*)) nicht hinreichend gut bewältigt werden. Wenn nun diese Anforderungen für das Lösen von Testaufgaben erforderlich sind, führt dies zu höheren empirischen Schwierigkeiten als sie für das eigentlich angenommene Niveau der Aufgabe (z. B. B1) erwartet werden. Nach dem Standard Setting würden solche Aufgaben auf einem höheren Niveau als B1 verortet, wohingegen jedoch die *sprachlich-kommunikativen* Anforderungen dieser Aufgaben auf B1 zu verorten sind. Dieses Paradox gilt auch für den umgekehrten Fall: Lernende könnten bestimmte Aufgaben weitaus häufiger lösen als auf einem bestimmten Kompetenzniveau erwartet, etwa weil bestimmte Aufgabenanforderungen Teil des Lehrplanes sind und besonders gefördert werden. Dies könnte zu empirisch geringeren Schwierigkeiten solcher Testaufgaben führen als erwartet.

Bei solchen Testaufgaben, deren empirische Schwierigkeiten stark von den a-priori Einschätzungen und den Expertenurteilen beim Standard Setting abweichen, ist Vorsicht geboten. Belässt man sie im Aufgabenpool, werden sie fälschlicherweise auf einem zu hohen oder zu niedrigen Niveau eingestuft und verzerren die Testwertinterpretation. Sinnvoll wäre es, solche Aufgaben aus dem Testpool herauszunehmen und didaktisch analysieren. Aus solch einer Analyse ließen sich dann wertvolle Anregungen für die Unterrichtsentwicklung gewinnen, etwa um Anforderungsbereiche zu analysieren, deren Beherrschung sinnvoll und wichtig wäre, die aber noch nicht ausreichend beherrscht werden. Die analysierten Testaufgaben wären eine gute Ausgangsbasis, um Lernaufgaben zur Entwicklung dieser Anforderungsbereiche zu erstellen.

4. Fazit

Zur Bildung von Kompetenzniveaus gibt es eine Vielzahl von Vorgehensweisen, die durch unterschiedliche Vorkenntnisse und den unterschiedlichen Einbezug von vorab formulierten Annahmen über die zu messende Kompetenz gekennzeichnet sind. Je nach Kontext sollten Verfahren gewählt werden, die so viel Vorwissen wie möglich nutzen. Wir haben drei verschiedene Vorgehensweisen vorgestellt; allen gemein ist die Bildung einer Kompetenzskala, auf der Aufgabenschwierigkeiten und Schülerkompetenzen zugleich dargestellt werden können. Dabei spielt die IRT-Skalierung eine zentrale Rolle, da sie diese gemeinsame Skalierung ermöglicht. Ist die Kompetenzskala dann in Kompetenzniveaus unterteilt, müssen diese inhaltlich mit Bezug auf die von unterschiedlich kompetenten Personen bewältigten Anforderungen beschrieben werden. Diese Beschreibung kann, wie oben ausführlicher dargestellt, theoriegeleitet oder empiriegestützt erfolgen, und sie kann entweder a priori oder post-hoc erfolgen, je nach Forschungsstand und Testkontext. Die Kompetenzniveaubeschreibung ermöglicht es, qualitative Testanforderungen und Testergebnisse in Form von Könnensbeschreibungen zu kommunizieren, zusätzlich zu rein quantitativen Testwerten. Somit können Testrückmeldungen nicht nur einer breiteren Öffentlichkeit verständlich kommunizieren werden, sie sind zudem informativer für pädagogisches Handeln. Beispielsweise können Testergebnisse gezielt daraufhin analysiert werden, welche Aufgabenanforderungen und Kompetenzen in einer bestimmten Klasse noch nicht ausreichend beherrscht werden; diese Anforderungsbereiche können im Unterricht aufgegriffen werden, wobei die oben genannten Aufgabenspezifikationen helfen können, Aufgaben einzusetzen, die die analysierten Kompetenzbereiche umsetzen. Nicht nur für den fremdsprachlichen Unterricht, auch für die Fremdsprachenforschung hat die Kompetenzniveaubeschreibung durch Aufgabenmerkmale Implikationen. So können solche Aufgabenmerkmale bei der Analyse von Aufgabenschwierigkeiten, bei der Konstruktion neuer Aufgaben und deren Konstruktvalidierung oder bei der Bestimmung von Kompetenzniveaus durch Expertenurteile helfen.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

Literaturverzeichnis

Alderson, J. Charles (1991), Bands and scores. In: Alderson, J. Charles & North, Brian (Eds.) (1991), *Language Testing in the 1990s*. London: Macmillan.

Alderson, J. Charles; Figueras, Neus; Kuijper, Henk; Nold, Günter; Takala, Sauli & Tardieu, Claire (2006), Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly* 3, 3-30.

ALTE Members (2007), *The CEFR Grid for Writing Tasks* v. 3.1 (analysis).* [Online: http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_1_analysis.doc.]

Beaton, Albert E. & Allen, Nancy L. (1992). Interpreting scales through scale anchoring. Journal *of Educational Statistics*, 17, 191-204.

Cizek, Gregory J.; Bunch, Michael B. & Koons, Heather (2004), Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice* 23: 4, 31-50.

Council of Europe (2001), *A Common European Framework of Reference for Language Learning and Teaching*. Cambridge: Cambridge University Press.

Council of Europe (2009), *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). A Manual*. Strasbourg: Language Policy Division.

Goldhammer, Frank & Hartig, Johannes (2007), Testwertinterpretation. In: Moosbrugger, Helfried & Kelava, Augustin (Hrsg.), *Test- und Fragebogenkonstruktion*. Berlin: Springer.

Harsch, Claudia (2007), *Der gemeinsame europäische Referenzrahmen für Sprachen. Leistung und Grenzen.* Saarbrücken: VDM.

Harsch, Claudia & Rupp, André (2011), Designing and scaling level-specific writing tasks in alignment with the CEFR: a test-centered approach. *Language Assessment Quarterly* 8: 1, 1-34.

Harsch, Claudia & Tiffin-Richards, Simon P. (2010), Setting standards in line with the Common European Framework of Reference. In: Harsch, Claudia; Pant, Hans A. & Köller, Olaf (Eds.) (2010), *Calibrating Standards-based Assessment Tasks for English as a First Foreign Language. Standard-setting Procedures in Germany*. Münster: Waxmann.

Hartig, Johannes (2007), Skalierung und Definition von Kompetenzniveaus. In: Beck, Bärbel & Klieme, Eckhart (Eds.) (2007), *Sprachliche Kompetenzen. Konzepte und Messung*. Weinheim: Beltz.

Helmke, Andreas & Hosenfeld, Ingmar (2004), Vergleichsarbeiten – Standards – Kompetenzstufen: Begriffliche Klärungen und Perspektiven. In: Jäger, Reinhold S. & Frey, Andreas (Eds.) (2004), *Lernprozesse, Lernumgebung und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert.* Landau: Verlag Empirische Pädagogik.

KMK (2003), *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Abschluss*. Darmstadt: Luchterhand.

KMK (2004), *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Hauptschulabschluss.* Darmstadt: Luchterhand.

Leucht, Michael; Harsch, Claudia & Köller, Olaf (in Revision), *Schwierigkeitsgenerierende Merkmale von Items zum Lese- und Hörverstehen im Fach Englisch.*

Mitzel, Howard C.; Lewis, Daniel M.; Patz, Richard J. & Green, Donald R. (2001), The Bookmark procedure: Psychological perspectives. In: Cizek, Gregory R. (Ed.) (2001), *Setting Performance Standards: Concepts, Methods and Perspectives*. Mahwah, NJ: Erlbaum.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.

Moosbrugger, Helfried (2007), Item-Response-Theorie (IRT). In: Moosbrugger, Helfried & Kelava, Augustin (Eds.) (2007), *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer Verlag.

Naumann, Johannes; Artelt, Cordula; Schneider, Wolfgang & Stanat, Petra (2010), Lesekompetenz von PISA 2000 bis PISA 2009. In: Klieme, Eckhardt; Artelt, Cordula; Hartig, Johannes; Jude, Nina; Köller, Olaf; Prenzel, Manfred; Schneider, Wolfgang & Stanat, Petra (Hrsg.) (2010), *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster: Waxmann.

Nold, Günter; Rossa, Henning & Chatzivassiliadou, Kyriaki (2008), Leseverstehen Englisch. In: DESI-Konsortium (Hrsg.) (2008), *Unterricht und Kompetenzerwerb in Deutsch und Englisch.* Weinheim: Beltz.

OECD (2009), *PISA 2009 Assessment Framework. Key competencies in reading, mathematics and science.* [Online: http://www.oecd.org/dataoecd/11/40/44455820.pdf.]

Rauch, Dominique & Hartig, Johannes (2007), Interpretation von Testwerten in der IRT. In: Moosbrugger, Helfried & Kelava, Augustin (Eds.) (2007), *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer Verlag.

Rasch, Georg (1960), *Probabilistic models for some intelligence and attainment tests.* Kopenhagen: The Danish Institute for Educational Research.

Rupp, André & Porsch, Raphaela (2010), Standard-setting item pool. In: Harsch, Claudia; Pant, Hans A. & Köller, Olaf (Eds.) (2010), *Calibrating Standards-based Assessment Tasks for English as a First Foreign Language. Standard-setting Procedures in Germany*. Münster: Waxmann.

Rupp, André; Vock, Miriam; Harsch, Claudia & Köller, Olaf (2008), *Developing Standards-based Assessment Tasks for English as a First Foreign Language - Context, Processes and Outcomes in Germany*. Münster: Waxmann.

Tiffin-Richards, Simon P. (2010), The Bookmark standard-setting Method. In: Harsch, Claudia; Pant, Hans A. & Köller, Olaf (Eds.) (2010), *Calibrating Standards-based Assessment Tasks for English as a First Foreign Language. Standard-setting Procedures in Germany*. Münster: Waxmann.

Weir, Cyril J. (2005), Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing* 22: 3, 281-300.

Zieky, Michael & Perie, Marianne (2006), *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service.

Anmerkungen

[1] „Kompetenzniveau“ und „Kompetenzstufe“ werden in diesem Beitrag synonym gebraucht. Da der Begriff „Stufe“ potenziell irreführend ist (z. B. Helmke & Hosenfeld 2004), verwenden wir, außer bei direkten Bezügen auf Studien, in denen der Begriff „Kompetenzstufe“ verwendet wird (z. B. PISA) durchgehend den Begriff „Kompetenzniveau“.

Claudia Harsch & Johannes Hartig (2011), Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 16: 2, 6-17. Abrufbar unter http://zif.spz.tu-darmstadt.de/jg-16-2/beitrag/Harsch_Hartig.pdf.