Ulrike Behrens, Sebastian Weirich

It takes a whole village… Interdisziplinäre Perspektiven bei der Entwicklung eines Testverfahrens zum Zuhören

Abstract The article shows how the perspectives of linguistics, literature, didactics, psychometrics, acting, administration and technology are interlocked in the construction of test instruments. The project *stim·mig* has developed and evaluated a new kind of test items for the assessment of listening competency in primary school. Principles of the item construction are being explained and illustrated on the basis of sample items. It becomes clear that interdisciplinary cooperation is indispensable for meeting the state of the art in the various aspects of test development.

1 Einleitung

Am Beispiel einer einzelnen Testaufgabe zum Hörverstehen bei Grundschulkindern soll gezeigt werden, wie die Perspektiven verschiedener Fachrichtungen und Teildisziplinen im Rahmen der Itementwicklung ineinandergreifen. Besonderes Gewicht kommt einerseits fachwissenschaftlichen und fachdidaktischen Fragen zu, die sich verschränken mit einer psychometrischen Perspektive. Im Fall von Hörverstehensaufgaben sind andererseits die Interpretation und stimmliche Gestaltung von Texten zentral. Es wird deutlich, dass interdisziplinäre Zusammenarbeit unabdingbar ist, wenn man den Anspruch hat, dass möglichst alle Aspekte der Aufgabenentwicklung dem jeweiligen *state of the art* entsprechen.

Die Zusammensetzung der Projektgruppe kann als interdisziplinär (und zudem international) bezeichnet werden:

- Ulrike Behrens ist Diplompädagogin und wissenschaftliche Mitarbeiterin am Institut für Germanistik der Universität Duisburg-Essen,
- Felix Bertschin ist ausgebildeter Schauspieler und Primarschullehrer sowie Dozent für Theaterpädagogik und Sprecherziehung an der FHNW Liestal,
- Ursula Käser-Leisibach arbeitet als Dozentin für Sprache, Sprachentwicklung und Kommunikation im Institut *Vorschul- und Unterstufe* der FHNW Brugg-Windisch,
- Michael Krelle hat Germanistik studiert und ist akademischer Rat im Bereich Sprachdidaktik an der Universität Paderborn,

- Sebastian Weirich ist als Diplom-Psychologe am Institut für Qualitätsentwicklung im Bildungswesen für die psychometrische Seite der Projekte im Bereich Grundschule zuständig,
- Claudia Zingg Stamm ist Dozentin an der Professur für Deutschdidaktik und ihre Disziplinen im Institut *Primarstufe* der FHNW Liestal.

Es wird sich im Laufe des Artikels zeigen, dass das noch nicht alle Perspektiven sind, die im Prozess der Aufgabenentwicklung einzunehmen waren und diesen teils erheblich beeinflussten. So waren z. B. aufseiten des IQB als Kooperationspartner für die Testdurchführung im Rahmen der VERA-Pilotierung zahlreiche technisch-administrative Aufgaben zu erbringen, die unmittelbar mit inhaltlichen Fragen der Testentwicklung im Zusammenhang standen. Dazu gehört u. a. auch die teilweise äußerst schwierige Klärung urheberrechtlicher Fragen.

2 Rahmenbedingungen

Die Einführung der Bildungsstandards im deutschsprachigen Raum brachte für die Deutschdidaktik nicht nur eine Neuausrichtung von Lehr- und Lernzielen im Sinne einer Kompetenzorientierung mit sich; Fachdidaktikerinnen und Fachdidaktiker waren gleichzeitig erstmals mit der Anforderung konfrontiert, den Erwerb von Kompetenzen und das Erreichen von Standards im Rahmen größerer Erhebungen messbar zu machen. Dies stellte für den Bereich mündlicher Leistungen eine besondere Herausforderung dar: Zum einen lassen sich produktive mündliche Leistungen schon aus technischen Gründen nicht im Rahmen etwa der Bildungsstandards-Normierung, der Ländervergleiche oder VERA-Pilotierungen erheben, denn diese wurden durchweg in Form von *paper and pencil*-Gruppentestungen durchgeführt. Zum anderen hielten sich die Erfahrungen mit der Messung von Zuhörkompetenzen insgesamt in Grenzen. Erste Ansätze einer Modellierung der Zuhörkompetenz und ihrer Testung waren inspiriert

- von Arbeiten aus dem englischsprachigen Raum (hier z. B. Buck 2001),
- von Vorgehensweisen im Bereich der Fremdsprachdidaktik (vgl. etwa Grotjahn 2000; Nold/Rossa 2007) sowie
- von Formaten der Lesetestung, wie sie seit PISA auch in Deutschland geläufiger wurden (vgl. Baumert et al. 2001).

Insbesondere letzterer Umstand führte dazu, dass die resultierenden Kompetenzstufenmodelle für Lesen und Zuhören erhebliche Parallelen aufweisen, was ihnen gelegentlich die Kritik eintrug, hier werde nicht Zuhörfähigkeit, sondern eigentlich v.a. modalitäts*unabhängig* Textverstehen modelliert:

> Tatsächlich weisen die Daten der Pilotierungsstudie zur Evaluation der Bildungsstandards für das Fach Deutsch in der Grundschule eine Korrelation auf latenter Ebene von .74 zwischen den Kompetenzbereichen Lesen und Zuhören auf (vgl. Behrens, Böhme u. Krelle 2009). Man findet also im Textverstehen offenbar eine grundlegende Teilkompetenz als Komponente sowohl der Lese- als auch der Zuhörkompetenz. Die naheliegende Kritik, dass mit ‚Textverstehen' weder wirklich Zuhörspezifisches noch gar Zuhörtypisches erfasst wird, ist demnach teilweise berechtigt. (Behrens 2010, S. 37)

Zwar gab es Versuche, beispielsweise auch die Nutzung paraverbaler Informationen für den Verstehensprozess in Testaufgaben zu überprüfen. Die Ergebnisse waren aber zunächst empirisch nicht robust genug, um die entsprechenden Fähigkeiten auf Schülerseite sinnvoll in die Kompetenzmodelle zu integrieren. Aus Sicht der Mündlichkeitsdidaktik besteht jedoch kein Zweifel, dass gerade diese Fähigkeiten spezifisch für das Zuhören sind.

3 Projekt *stim·mig*

Die so umrissene Situation ist der Ausgangspunkt für das deutsch-schweizerische Projekt *stim·mig*, um das es im Folgenden gehen soll: Ziel war die Entwicklung und Evaluation von Zuhöraufgaben, die neben Verständnisfragen auf der verbalen Ebene auch Items enthalten, die sich ausschließlich auf Basis paraverbaler Informationen lösen lassen.

Die im Rahmen des Projekts entwickelten Aufgaben beziehen sich teils auf literarische Texte, teils auf Radiosendungen zu Sachthemen. Zusätzlich wurden Items entwickelt, die sich nicht auf einen längeren Aufgabenstamm beziehen. Hier müssen z. B. bestimmte Klangmerkmale (beim Flüstern, bei Heiserkeit, beim Kauen etc.) oder die pragmatische Funktion einer Betonung erkannt werden, oder die Testteilnehmer beurteilen authentische Vorleseleistungen von Grundschulkindern. Die Aufgaben wurden im Rahmen der Pilotierung der Vergleichsarbeiten in der dritten Jahrgangsstufe (VERA-3) im Frühjahr 2015 an einer Stichprobe von 5796 Schülerinnen und Schülern evaluiert.

Aufgrund der insgesamt großen Aufgabenmenge, die eine Schülerin oder ein Schüler allein nicht bewältigen könnte, bearbeitet jedes Kind dabei immer nur einen Teil aller Aufgaben. Die Aufgaben werden in Blöcken gruppiert, wobei jeder Aufgabenblock eine unterstellte Bearbeitungszeit von 20 Minuten hat. Jedes Testheft, das ein Kind bearbeitet, besteht dabei aus insgesamt vier Aufgabenblöcken, benötigt also eine planmäßige Bearbeitungszeit von $4 \cdot 20 = 80$ Minuten. Die Systematik, nach der jeweils vier Aufgabenblöcke für ein Testheft zusammengestellt werden, wird durch das *Testdesign* definiert.

Im Rahmen der Studie wurden die Aufgaben des *stim·mig*-Projekts in das VERA-Testdesign integriert; jedes Testheft konnte also sowohl Zuhöraufgaben aus *stim·mig* als auch Lese- bzw. Zuhöraufgaben aus VERA enthalten. Auf diese Weise wurden insgesamt 56 unterschiedliche Testhefte zusammengestellt, wobei sich einige Testhefte in der Auswahl der Aufgabenblöcke teilweise (nicht vollständig) überschneiden. Für jede der an der Studie teilnehmenden Klassen wurde zufällig ein Testheft zur Bearbeitung ausgewählt.

Um eine hohe Durchführungsobjektivität zu gewährleisten, wurde die Studie von instruierten Testleitern durchgeführt. Bei den meisten Zuhöraufgaben wird der gesamten Klasse zunächst ein Text oder Textausschnitt von einer CD vorgespielt, der in seiner Länge zwischen vier und zehn Minuten variiert. Anschließend sind die Kinder aufgefordert, in Stillarbeit zu diesem Text einige Fragen im Testheft zu beantworten. Die Fragen sind so konstruiert, dass eine richtige Beantwortung ein Verständnis des zuvor gehörten Textes bzw. ein korrektes Dekodieren paraverbaler Merkmale voraussetzt.

Nach der Testdurchführung werden die Testhefte wieder eingesammelt und die Antworten der Kinder in einen Datensatz übertragen, wobei jede Antwort eines jeden Kindes wahlweise als 0 (falsch) oder 1 (richtig) kodiert wird. Damit liegt für jedes Item eine Reihe von Itemantworten vor, aus der sich etwa die relative Schwierigkeit des Items (bzw. seine Lösungshäufigkeit) bestimmen lässt.

Insgesamt nahmen an der gemeinsamen Studie (VERA und *stim·mig*) 5796 Kinder in 299 Klassen teil. Von ihnen bearbeiteten 4158 Kinder in 210 Klassen Aufgaben des *stim·mig*-Projekts.

Für die Auswertung kommen – wie bei vergleichbaren Schulleistungsstudien üblich – Modelle der probabilistischen Testtheorie (oder Item Response Theory) infrage. Diese Modelle unterstellen eine latente, nicht direkt zu beobachtende Personenfähigkeit (hier: Zuhörkompetenz). Je höher diese Fähigkeit für ein Kind ist, desto höher sollte die Wahrscheinlichkeit sein, dass es eine bestimmte Testaufgabe korrekt löst. Umgekehrt kann für jede Testaufgabe ein Schwierigkeitswert (oder Schwierigkeitsparameter) bestimmt werden: Je größer die Schwierigkeit einer Aufgabe, desto geringer ist die Wahrscheinlichkeit, dass ein bestimmtes Kind diese Aufgabe korrekt bearbeitet. In weiterführenden Analysen kann nun versucht werden, die empirisch ermittelten Schwierigkeiten der Testaufgaben auf ihre spezifischen Eigenschaften zurückzuführen.

Eine erste Sichtung der Itemparameter ergab insgesamt zufriedenstellende Werte, allerdings variieren die Schwierigkeiten der Aufgaben zum Teil erheblich. Über Zusammenhänge mit verschiedenen möglichen Merkmalen der Aufgaben

und Items kann beim jetzigen Stand der Datenauswertung noch nichts gesagt werden. Ein ausführlicher Projektbericht mit detaillierten Analysen ist in Planung.

In der intensiven Projektarbeit zeigte sich: Die Entwicklung von Zuhöraufgaben ist ein äußerst komplexes Unterfangen, das allen Kooperationspartnern wiederholte Perspektivenwechsel und einen vielfach verschränkten interdisziplinären Diskussionsprozess abverlangt. Die Produktion und gemeinsame Diskussion der Aufgaben erforderte insgesamt vier Projekttreffen und etwa zwanzig Onlinekonferenzen. Im Folgenden sollen die verschiedenen Perspektiven dargestellt werden, die in den verschiedenen Stadien eines solchen Konstruktionsprozesses zu berücksichtigen sind. Für die Aufgabenentwicklung ist das von beträchtlicher Relevanz, weil ihre Berücksichtigung nicht nur den Prozess bereichert, sondern in jedem einzelnen Fall auch dazu führen kann, dass eine in anderen Hinsichten gelungene Aufgabe sich als nicht brauchbar erweist. Im Laufe der Entwicklungsphase wurde etwa die Hälfte der Aufgabenvorschläge in verschiedenen Stadien wieder verworfen. Sehr häufig sind zudem abwägende Entscheidungen zwischen unterschiedlichen Interessen oder Sichtweisen zu treffen. Das betrifft nicht nur den hier exemplarisch nachgezeichneten Entwicklungsprozess, sondern die (Test-) Aufgabenkonstruktion generell – ein Umstand, der zu bedenken ist, wenn Testaufgaben einer kritischen Würdigung durch Dritte unterzogen werden.

4 Die Beispielaufgabe: Aufregung im Schloss

Im Folgenden soll anhand eines konkreten Aufgabenbeispiels verdeutlicht werden, wie verschiedene fachliche Perspektiven die Entscheidungen im Verlauf der Aufgabenentwicklung nicht nur beeinflussen, sondern überhaupt erst ermöglichen. Als Beispiel zur Illustration wurde hier die Aufgabe *Aufregung im Schloss* gewählt. Sie bezieht sich auf einen Auszug aus dem Kinderbuch *Die wilde Sophie* des Schweizer Schriftstellers Lukas Hartmann. Für die Testaufgabe wurde der Beginn des sechsten Kapitels ausgewählt, als Lesung inszeniert und in Hörbuchform produziert. Der Sprecher Felix Bertschin hatte dabei die Aufgabe, den Figuren unterschiedliche Stimmen zu verleihen, die zur jeweiligen Rolle passen. Entsprechend der auch auf der verbalen Textebene deutlich überzeichneten Figurencharakteristika (die verängstigten Diener, der polternd-autoritäre König, die bodenständige Köchin etc.) sind die verschiedenen Stimmen in der Lesung gut zu unterscheiden.

Der ausgewählte Ausschnitt umfasst 700 Wörter und dauert in seiner Audiofassung siebeneinhalb Minuten. Im Buch gehen ihm fünf Kapitel voraus, in denen das Setting und die Figuren der Geschichte eingeführt werden. Im Mittelpunkt steht Prinz Jan, der vor allem von seinem Vater, König Ferdinand, aus Angst und Sorge um sein Wohlergehen in übertrieben-skurriler Weise behütet und umsorgt

wird. Ein ganzes Heer von Angestellten, darunter ein „Nebenhergeher", ein „Insektenjäger", ein „Kleideranwärmer" und ein „Lebertranverwalter", kümmern sich um den Prinzen, der unter diesen Bedingungen zu einem schüchternen und ängstlichen Jungen heranwächst. Dann aber lernt er Sophie kennen, die Titelfigur des Buches, die ganz anders lebt als er und in großer Selbstverständlichkeit all die Dinge tut, die Jan verboten sind.

In dieser Situation beginnt das sechste Kapitel und damit der Textausschnitt, der als Aufgabenstamm für die Testaufgabe fungiert: Die beiden Diener Raimund und Stanislaus wachen auf und entdecken, dass der Prinz, auf den sie offenbar hätten aufpassen sollen, aus seinem Bett verschwunden ist. Noch während sie panisch versuchen zu entscheiden, was zu tun ist, taucht der König auf und will wissen, was los ist. Unter Tränen gestehen sie das Verschwinden ihres Schutzbefohlenen und lösen den befürchteten Wutanfall aus. Der König droht ihnen mit Kerkerhaft und lässt durch zehn Gongschläge Großalarm auslösen, weil er ein Verbrechen befürchtet. Der pragmatische Einwand der Köchin, der Prinz könne freiwillig weggelaufen sein, aus Spaß oder weil er Strafe für ein Vergehen befürchte, wird abgeschmettert, denn dazu habe der Prinz keinerlei Grund. Die angeordnete Durchsuchung des Schlosses ergibt schließlich ein aufgebrochenes Fenster und Blutspuren – für den König der Beweis für eine Entführung. Die Diener entdecken jedoch weitere Hinweise auf die Theorie der Köchin: Die meisten Glasscherben liegen draußen, und ein Baldrianfläschchen ist leer. Der König ist außer sich vor Empörung, dass sein Sohn den Dienern Schlafmittel in den Tee geschüttet haben soll, um ihnen zu entwischen. Er befiehlt den verängstigten Dienern aufgebracht, den Prinzen lebendig zurückzubringen, während er selber einem Zusammenbruch nahe ist.

Hier endet der Textausschnitt. Es folgen insgesamt zehn Testfragen. Diese beziehen sich zur Hälfte auf Inhalte des Textes auf der verbalen Ebene und entsprechen in ihrer Konstruktionsweise den Testaufgaben, wie sie bislang in Normierungsstudien und Vergleichsarbeiten eingesetzt wurden (vgl. auch Behrens et al. 2009). Dies entspricht zum einen der mutmaßlichen Zuhörhaltung der Kinder, die ja zunächst einmal die Geschichte und den Verlauf der Ereignisse verfolgen. Zum anderen gewinnt man so Hinweise darauf, ob die gewählten Texte und neu entwickelten Aufgaben hinsichtlich ihrer Schwierigkeit denen entsprechen, auf deren Basis die aktuellen Modelle konstruiert sind. Diese „verbalen" Items dienen also der Verknüpfung mit bisherigen Modellen und damit auch der Validierung der Aufgabe. Um diese Vergleichbarkeit zu erreichen, sollen die Items nach Möglichkeit auf unterschiedliche Ebenen des Textverstehens entsprechend den fünf Niveaus des Kompetenzstufenmodells zum Hörverstehen abzielen (vgl. IQB 2013, S. 7–10).

Als einfach (Stufe 1) gelten dabei Aufgaben zum Wiedererkennen prominenter Einzelinformationen. Anders als bei Lesetexten bezieht sich bei Hörverstehensaufgaben *prominent* allerdings nicht unbedingt auf die Platzierung beispielsweise zu Beginn eines Textes oder in dessen Titel, wo eine gesuchte Information leicht (wieder) aufzufinden ist. Im akustischen Modus könnte beispielsweise eine Information vom Textanfang im Gedächtnis schon von jüngeren Informationen überlagert sein, wenn am Ende des Hörens Fragen dazu zu beantworten sind. Eine größere Rolle spielt möglicherweise die zeitliche Nähe zum Höhepunkt einer Erzählung, mehrfache Erwähnung im Text, eine besondere akustische Hervorhebung o.ä. Diese theoretischen Überlegungen ergeben sich aus psychologischen bzw. sprachdidaktischen Arbeiten zum Zuhörprozess (z. B. Imhof 2003). Eine empirisch gestützte Modellierung steht aber hier bislang noch aus. Auf den höheren Kompetenzstufen sind aus mehreren Informationen Schlüsse zu ziehen; Items auf Stufe 5 sind u. a. gekennzeichnet durch eigenständigere Beurteilungsleistungen. Die Items zu *Aufregung im Schloss* erweisen sich allerdings (zumindest für die Kinder der Pilotierungsstichprobe) mit Lösungshäufigkeiten zwischen 50% und 90% als durchgängig eher leicht – die schwierigsten vier Items dieser Aufgabe rangieren auf den Kompetenzstufen 2 und 3.

Das folgende Beispielitem wird von 84% der Kinder richtig gelöst und ist somit der Kompetenzstufe 1 zuzuordnen:

Wer sind Raimund und Stanislaus?

☐ A) Zwei Diener
☐ B) Zwei Soldaten
☐ C) Zwei Einbrecher
☐ D) Zwei Prinzen

Die so charakterisierten *verbalen* Items wurden den Kindern wie bislang üblich in schriftlicher Form vorgelegt; sie mussten also nach dem Hören des Textes gelesen und im Testheft beantwortet werden. Das Design der Gesamtstudie sieht jedoch vor, dass die Items den Kindern teilweise zusätzlich akustisch präsentiert werden. Dadurch sollen die Projektergebnisse unter anderem zu einer weiteren Klärung des Verhältnisses von Lese- und Zuhörfähigkeiten beitragen.

In einem zweiten Teil werden den Kindern Fragen gestellt, die sich auf *prosodische* bzw. *paraverbale* Merkmale von Textteilen beziehen. Sie bilden den innovativen Kern des Projekts. Solche Items sind zuerst im Projekt *ohrwärts* (vgl. Zingg Stamm et al. 2014) als Element eines Lehrmittels zur Diagnose und Förderung der Hörkompetenzen in der Grundschule entwickelt und erprobt worden, dort allerdings zunächst im Rahmen einer einzelnen Aufgabe (d. h. zu einem einzigen Hörtext). Effekte des spezifischen Textes konnten so zunächst nicht überprüft

werden; auch konnten die neuen Aufgabenformate nur in moderatem Umfang (N=200) systematisch evaluiert werden. Andererseits erlaubte die vergleichsweise geringe Zahl an Testpersonen, im Rahmen von Beobachtungen in den Versuchsklassen das praktische Funktionieren des neuen Formats und die Einschätzung der Kinder einzuholen. Jenseits von psychometrischen Zugangsweisen kommt hier schulpädagogische Expertise zum Tragen, die in der kaum normierbaren, komplexen Situation im Klassenzimmer in der Lage ist, die je individuellen Umgangsweisen der Kinder mit den verschiedenen Herausforderungen des Tests einzuschätzen. Dieser Weg ist, wiewohl wenig standardisiert, vermutlich der einzige, auf dem man mögliche Störungen wie unerwartete Heiterkeit an bestimmten Textstellen, aber auch motivationale Aspekte wie Langeweile, Überforderung oder andere Irritationen erheben kann.

Das Projekt *stɪm·mig* zeichnet sich demgegenüber dadurch aus, dass erstmals verlässliche Itemkennwerte generiert werden: Es werden zum einen in zwölf zusätzlich entwickelten Aufgaben Texte und Genres systematisch variiert und diese zum anderen an einer großen Stichprobe evaluiert.

Typisch für diese *paraverbalen* Items des jeweils zweiten Testteils ist, dass die Unterschiede zwischen den Antwortoptionen auf der verbalen Ebene nicht zu erkennen sind. Die richtige Lösung kann nur gefunden werden, indem man prosodische Informationen auswertet. Auch hier wurde ein Teil der Items in den zwei Varianten *nur gehört* (Variante 1) und *gelesen und gehört* (Variante 2) evaluiert (Näheres zum Design s.u.). Hierzu ein Beispielitem:

Die Königin ist leicht genervt, weil der König so laut ist. Wo hörst du das am besten?

Variante 1:	Variante 2:	
☐ A)	☐ A)	Weshalb weckst du mich mitten in der Nacht?
☐ B)	☐ B)	Weshalb weckst du mich mitten in der Nacht?
☐ C)	☐ C)	Weshalb weckst du mich mitten in der Nacht?
☐ D)	☐ D)	Weshalb weckst du mich mitten in der Nacht?

Diese Frage „Weshalb weckst du mich mitten in der Nacht?“ kann neutral-interessiert, empört oder auch verzweifelt gemeint sein – entscheidend hierfür ist die stimmliche Gestaltung, die von den Testpersonen richtig zugeordnet werden muss. Alle Items des zweiten Aufgabenteils folgen dieser Logik. Das hat Konsequenzen für die Modalität der Darbietung: Diese *paraverbalen* Items können nur dann gelöst werden, wenn sie den Kindern zumindest *auch* akustisch dargeboten werden.

5 Die Textauswahl

Eine der ersten Herausforderungen stellt die Auswahl geeigneter Texte dar. Sie sollen den Interessen der Zielgruppe (hier: von Grundschulkindern) entgegenkommen. In sprachlicher wie in inhaltlicher Hinsicht muss ein Text so gewählt werden, dass er die Verstehensfähigkeiten der Kinder weder übersteigt noch sie unterfordert. Dabei macht es einen Unterschied, ob der Text von den Kindern gelesen oder gehört werden soll. So könnten im Beispieltext etwa Wörter wie *steifbeinig, totenblass, Thronsaal* für junge Leser erhebliche Hürden darstellen, Begriffe wie *Leibarzt, einschläfern, Baldrian* wären ggf. zu erläutern. Wenn diese Wörter jedoch nicht selbst erlesen werden müssen, dann gelingt Kindern vermutlich das Verstehen, also die Rekonstruktion von Bedeutung, leichter, da sie kognitiv von den Leseanforderungen entlastet sind. Die stimmliche Gestaltung durch den Sprecher kann dies zusätzlich unterstützen, indem wichtige Stellen besonders hervorgehoben werden oder die herausgehobene Bedeutung einer Phrase (wenn etwa der König sagt: *Das ist eine himmelschreiende Verleumdung!*) stimmlich markiert wird. Karla Müller schreibt,

> dass das Verstehen von Figurenperspektiven mitunter auditiv besser gelingt als beim Lesen. Indem Unbestimmtheitsstellen des Textes durch prosodische Elemente gefüllt werden, kann der emotionale Zustand einer Figur in einer Hörfassung leichter fasslich werden. (Müller 2012, S. 76)

Ein weiteres Kriterium für die Texteignung ergab sich aus den Anforderungen der Prosodie-Items: Es war darauf zu achten, dass im Text gesprochen wird, dass Dialoge vorkommen und möglichst auch verschiedene Emotionen geäußert werden. Andererseits durften nicht zu viele Figuren sprechen, damit jeder einzelnen von einem einzigen Sprecher ein charakteristisches Stimmprofil zugeordnet werden konnte. So konnten in einigen Items den Figuren weitere, im Text nicht enthaltene Äußerungen in den Mund gelegt werden – die Kinder hatten dann zu entscheiden, wer jeweils spricht.

Die Eignung eines Textes zu bestimmen, setzt also die Analyse der inhaltlichen und sprachlichen Potenziale *und* Herausforderungen voraus. Dies ist zunächst eine literaturwissenschaftliche Anforderung; sofern sich die Eignung auf eine bestimmte Altersgruppe bezieht, kommen zudem sprach- und literaturdidaktische Erwägungen in den Blick. Das ist auch der Fall bei der Auswahl von Texten etwa zur Behandlung im Unterricht. Wenn der Text zu *Test*zwecken eingesetzt werden soll, ergeben sich aus i. w. S. psychometrischer Perspektive weitere Anforderungen:

Zunächst ist zu beachten, dass nicht durch die Thematik einzelne Teilgruppen (z. B. Mädchen, Stadtkinder…) durch die Wahl der Texte systematisch bevorzugt

werden. Aus dem gleichen Grund ist es wünschenswert, dass ein Text möglichst nicht vielen Kindern (z. B. aus einer bestimmten Region oder auch Bildungsschicht) bereits bekannt ist. Das schließt zum einen (Serien von) Geschichten aus, die als Bücher oder auch in anderen medialen Adaptionen sehr populär sind. Zum anderen kommen aber auch keine Texte infrage, die in verbreiteten Lehrwerken für die Grundschule enthalten sind. Erfolg verspricht vor allem die Sichtung von neu erschienenen (Hör-)Büchern, aber auch von Lektürelisten aus dem (deutschsprachigen) Ausland, weil die Überschneidungen sich nach unseren Erfahrungen in Grenzen halten: Texte, die in der Schweiz sehr verbreitet rezipiert werden, können in Deutschland weitgehend unbekannt sein.

Eine sinnvolle Verwendung der gesamten Bearbeitungszeit erfordert Texte, die bei geringem zeitlichen Umfang möglichst viele sinnvolle Testitems ermöglichen. Das bedeutet, dass im Allgemeinen kürzere Texte längeren vorzuziehen sind, sofern sie inhaltlich ausreichend komplex sind, sodass sich Fragen auf verschiedenen Schwierigkeitsstufen konstruieren lassen. Diese Anforderung bezieht sich nicht nur auf das Kriterium der Testökonomie (d. h. mit möglichst wenig zeitlichem – und finanziellem – Aufwand möglichst viel Information über die Kompetenzen der Testteilnehmer gewinnen; vgl. Moosbrugger/Kelava 2007, S. 20f.), sondern berührt auch Fragen der Validität: Da beim Zuhören mit ansteigender Textlänge vermutlich auch der Anteil der Erinnerungsleistung und ggf. Konzentrationsfähigkeit am getesteten Konstrukt ansteigt, ist es theoretisch möglich, dass Testaufgaben auf Basis sehr langer Texte Anderes testen als solche, die sich auf kürzere Texte beziehen. In der Bildungsstandards-Normierung wurden deswegen Texte mit großem Umfang in mehreren Abschnitten präsentiert, auf die sich dann jeweils nur einige Testfragen bezogen. Der Effekt solcher Maßnahmen ist u. W. aber bislang noch nicht quantifiziert worden. Die im hier beschriebenen Projekt eingesetzten Texte sind nicht länger als zehn Minuten.

Hier war zusätzlich zu bedenken, dass die Aufgaben im Rahmen einer Großstudie eingesetzt werden sollten, die vereinheitlichte zeitliche Abläufe nötig machte: Damit im Rahmen der VERA-Pilotierung die Reihenfolge von Testblöcken zum Lesen und zum Zuhören variiert werden kann, muss die Gesamtbearbeitungszeit für jeden Block gleich lang sein (hier i. d. R. 20 Minuten; s. o.). Sprengt eine Aufgabe diesen zeitlichen Rahmen, so muss ggf. der Text sinnvoll gekürzt oder auf Items verzichtet werden. Dass aus urheberrechtlichen Gründen Kürzungen und Veränderungen innerhalb der Texte in der Regel nicht infrage kommen, schränkt die Auswahl an geeigneten Texten zusätzlich ein.

So kann es passieren, dass ein großer Pool an prinzipiell geeigneten Texten erheblich reduziert werden muss, weil einzelne Texte

- thematisch einzelne Gruppen von Kindern systematisch bevorzugen oder benachteiligen,
- Grundschulkinder beim Zuhören unterfordern würden,
- zu wenig Material für Items auf verschiedenen Schwierigkeitsniveaus bieten,
- zu lang sind und nicht sinnvoll gekürzt werden können.

6 Der Textausschnitt

Erfahrungsgemäß können viele (literarische) Texte die oben genannten Bedingungen nicht erfüllen. Je nach Gestaltung enthält eine Lesung pro Minute etwa 100 Wörter; der Beispieltext *Aufregung im Schloss* dauert in seiner Audiofassung ca. siebeneinhalb Minuten und entspricht damit in etwa dem angestrebten Textumfang. Selbst dieser Umfang wird in den unsystematisch erhobenen Protokollen, in denen die Testleiter besondere Beobachtungen aus den Testsitzungen notieren, gelegentlich als für die Kinder zu lang vermerkt. In sich abgeschlossene Texte dieser Länge finden sich aber zumeist für ein jüngeres Publikum, zum Beispiel in Form von Bilderbüchern oder etwa den Pixi-Büchern des Carlsen-Verlags. Typischerweise machen hier aber die Illustrationen einen beträchtlichen Anteil des Buchinhalts aus, der für den Textinhalt nicht ohne Weiteres verzichtbar ist. Zudem erweisen sich die Geschichten in ihrer Erzählstruktur häufig als zu eindimensional, sodass es schwerfällt, Fragen auf höheren Anforderungsniveaus zu formulieren.

Eine Alternative zu solchen abgeschlossenen Geschichten sind Auszüge aus längeren Texten. Häufig eignet sich nur der Beginn einer Erzählung, weil im späteren Verlauf auf Informationen aus der Exposition zurückgegriffen wird und der Ausschnitt ohne diese Informationen nicht verständlich ist. Im Textauszug, der hier zur Illustration dient, ergibt sich diese Schwierigkeit nur an den zwei Stellen, an denen König und Königin bei ihren Vornamen – Ferdinand und Isabella – genannt werden. Im einen Fall wird gleich im Folgesatz klar, um wen es sich handelt: „*Mein Sohn* ist entführt worden!“ Dass es sich bei Isabella um die Königin handelt, kann angenommen werden. Es ergibt sich nicht zwingend aus dem Text, ist aber für das Textverständnis auch nicht von großer Bedeutung. Insofern kann der Ausschnitt gehört werden wie der Beginn einer Geschichte, freilich mit einem offenen Ende, das auf eine notwendige Fortsetzung hinweist. Die Vorgeschichte, beginnend noch vor der ersehnten Geburt des Prinzen Jan, die in den vorausgehenden fünf Kapiteln dargestellt wird, ist für das Verständnis *dieses* Ausschnittes nicht erforderlich. Der Buchtitel *Die wilde Sophie* musste hingegen gegen eine passendere Überschrift ausgetauscht werden, um die Zuhörenden nicht zu verwirren.

Neben den Auswahlkriterien der passenden Textlänge und einer gewissen Abgeschlossenheit ist für erzählende Texte entscheidend, dass zentrale Strukturelemente von Geschichten enthalten sind. Hierzu zählt neben der angemessenen, ggf. sukzessiven Einführung von Setting und Personal eine Komplikation bzw. ein Planbruch als Voraussetzung der Erzählwürdigkeit (vgl. Quasthoff 1980, S. 57–67), im Falle des Beispiels das Verschwinden des Prinzen, das die weiteren Aktivitäten auslöst.

7 Die Itemkonstruktion

Die Normierungsstudie im Rahmen der Evaluation der Bildungsstandards führte für den Kompetenzbereich *Zuhören im Grundschulalter* zur Konstruktion eines fünfstufigen Kompetenzstufenmodells (vgl. IQB 2013). Die Beschreibungen der einzelnen Stufen enthalten neben den jeweiligen kognitiven Anforderungen auch Hinweise zum Format der Items, die auf der jeweiligen Stufe typischerweise gelöst werden. So werden auf der Stufe I „die Anforderungen […] vor allem dann bewältigt, wenn die Aufgaben ein Multiple-Choice-Format haben" (IQB 2013, S. 7); auf Stufe III werden „lenkende Hinweise, wie sie in Multiple-Choice-Aufgaben enthalten sind, […] dabei nicht mehr durchgängig benötigt, um die Anforderungen zu bewältigen" (IQB 2013, S. 9), und auf der Stufe V heißt es: „Die Schülerinnen und Schüler bewältigen die Anforderungen dabei meist auch im Rahmen von komplexen offenen Aufgaben" (IQB 2013, S. 10). Insgesamt zeigt sich, dass geschlossene Aufgabenformate dazu beitragen, Items leichter zu machen, während ein offenes Aufgabenformat eine Aufgabe tendenziell schwieriger macht. Um den Einfluss des Aufgabenformats zu reduzieren und von anderen Effekten abgrenzen zu können, wurde im Projekt durchgängig auf geschlossene Itemformate gesetzt. Dabei handelte es sich in aller Regel um Multiple-Choice-Aufgaben mit vier Antwortoptionen und genau einer richtigen Lösung. Gelegentlich wurden auch Wahr-Falsch-Aufgaben eingesetzt, bei denen jeweils vier zu bewertende Aussagen zu einem Item gebündelt werden, um die Ratewahrscheinlichkeit zu reduzieren.

Für eine Mischung aus geschlossenen und offenen Aufgaben spricht in der Regel neben einer abwechslungsreicheren Testgestaltung vor allem die Überlegung, dass nur in (halb-)offenen Aufgabenstellungen, also Aufgaben, in denen die Testteilnehmer selbst formulierte Antworten geben müssen, Wissen erhoben wird, das die Schülerinnen und Schüler eigenständig produzieren können (statt auf bloßes Wiedererkennen zu setzen). Oft wird auch argumentiert, dass komplexere kognitive Prozesse, mithin tiefergehende Verstehensleistungen im Multiple-Choice-Format nicht ermittelt werden können.

Andererseits hat der Einsatz von halboffenen und offenen Items den Nachteil, dass die Stärke des Effekts, den das Format auf die Lösungswahrscheinlichkeit hat, kaum bestimmt werden kann. Zudem gibt es administrative und auch finanzielle Konsequenzen: Während die Auswertung geschlossener Items sehr eindeutig mithilfe von Schablonen, unter Umständen sogar maschinell erfolgen kann, müssen von den Testteilnehmern selbst formulierte Lösungen von eigens geschulten Personen beurteilt werden. Sowohl die Erstellung und Erprobung der dafür erforderlichen Auswertungsanleitung als auch die Schulung der Kodiererinnen und Kodierer erfordert erhebliche zeitliche und personelle Ressourcen, die in der Gesamtplanung und -kalkulation entsprechender Projekte zu berücksichtigen sind. Das trifft insbesondere dann zu, wenn in einem größeren Projekt eine *Gruppe* von Kodierenden dahin geführt werden muss, Schülerlösungen gleichsinnig zu beurteilen. Eine zu geringe Übereinstimmung in den Kodierungen derselben Lösung weist auf unpräzise Auswertungskriterien hin, die selbst im Falle eines einzelnen Raters zu mangelnder Reliabilität der Testergebnisse führen könnte. Dies ist ein Beispiel für den oben angesprochenen Abwägungsprozess, der hier (neben den ökonomischen) zwei Aspekte von Validität betrifft: Offene Items können die Validität erhöhen, weil alltagsrelevantere Wissensarten in den Blick kommen. Andererseits kommen aber auch zusätzliche Kompetenzen (hier: Schreiben) ins Spiel, deren Einfluss auf das Lösungsverhalten nicht quantifizierbar ist.

Die Entscheidung für ein durchgängig geschlossenes Aufgabenformat, wie sie im hier vorgestellten Projekt getroffen wurde, macht es erforderlich, zu jeder Frage neben der richtigen Antwort drei plausible, aber eindeutig falsche Antworten (Distraktoren) zu formulieren. Inwiefern das gelungen ist, wird die Itemanalyse ergeben: Die Häufigkeit, mit der die Kinder ein Item richtig beantworten, gibt Auskunft über dessen Schwierigkeit; inwieweit die Distraktoren gut gewählt wurden, zeigt sich, wenn man das Antwortverhalten bei ungelösten Items betrachtet: Idealerweise sollten alle falschen Optionen etwa gleich häufig angekreuzt werden. Der Grund ist ein statistischer: Falls sich eine der falschen Antworten als abwegig erwiese und von kaum jemandem gewählt würde, blieben faktisch nur drei statt vier Optionen, was die Ratewahrscheinlichkeit von 25% auf 33% erhöhen und damit das Item im Vergleich deutlich leichter machen würde, ohne dass die erforderliche Zuhörleistung tatsächlich komplexer wäre.

8 Die Testfragen

Eine Geschichte zu verstehen bedeutet, auf Rezipientenseite eine mentale Repräsentation der Textinhalte zu konstruieren, die der Intention aufseiten des Autors entspricht (vgl. Schnotz 1996, S. 972). Für die Testkonstruktion ist zu bestimmen,

welches die für solches Verstehen relevanten Inhalte sind, was man also von kompetenten Rezipienten an Verstehensleistungen erwarten kann bzw. muss. Hier macht es nach sprachdidaktischen Einsichten in die Natur und Charakteristika mündlicher Kommunikation für die Testkonstruktion einen wesentlichen Unterschied, ob der Text gehört oder gelesen wird: Während Lesetestfragen durchaus auch auf eine eingehendere, zielgerichtete Re-Lektüre abzielen können, ist dies beim Zuhören nicht der Fall. Wegen der Flüchtigkeit akustischer Signale (vgl. Fiehler 2014; Fiehler et al. 2004) müssen Zuhöritems solche Informationen und Schlussfolgerungen adressieren, die Zuhörende typischerweise bereits während des Hörens in ihre mentale Repräsentation integrieren. So verbieten sich – um nur ein Beispiel zu nennen – Fragen nach ausschmückenden, aber inhaltlich nebensächlichen Details. Beim Entwurf von Testfragen ist es daher ratsam, sich Notizen zu wichtigen Textinhalten möglichst bereits nach dem *ersten Hören* zu machen bzw. Fragen, die auf Basis eines (schriftlichen) Transkripts formuliert wurden, in dieser Hinsicht kritisch prüfen zu lassen.

Im Unterschied zum auch empirisch gut erforschten Leseverstehen stecken Modelle zur Zuhörkompetenz nach wie vor in den Kinderschuhen. Unstrittig ist, dass Rezipienten beim Zuhören komplexe, multimodale Informationen auswerten, um Sinn zu rekonstruieren. Die bislang empirisch überprüften Modelle berücksichtigen dabei zu einem wesentlichen Anteil den Aspekt des Textverstehens. Vergleichende Untersuchungen von Leistungen in Lese- und Zuhörtestungen (vgl. Böhme et al. 2010) weisen auf erhebliche Überschneidungen beider Modalitäten hin, die derzeit am ehesten dadurch erklärt werden können, dass man eine modalitätsunabhängige bzw. -übergreifende Textverstehenskompetenz annimmt. Zusammenfassend schreibt Böhme,

> dass sowohl für das Hör- als auch für das Leseverstehen allgemeine Aufgabenmerkmale, die sich auf das Textverstehen beziehen, bei der Erklärung der Itemschwierigkeit einen wichtigen Stellenwert einnehmen. Um größere Teile der Varianz in der Itemschwierigkeit erklären zu können, sind jedoch Merkmale nötig, die spezifische Eigenschaften des Hör- beziehungsweise Leseverstehens beschreiben. Diese Befunde sind erwartungskonform, da in der theoretischen Betrachtung der beiden rezeptiven Sprachkompetenzen betont wird, dass die kognitiven Informationsverarbeitungsprozesse weitreichende Ähnlichkeiten aufweisen, auf der Ebene der Rezeption jedoch die verschiedenen Inputstimuli (visuell vs. auditiv) berücksichtigt werden müssen. (Böhme 2011, S. 92)

Zu den *zuhörspezifischen* Verstehensleistungen gehört ohne Zweifel die Bedeutungsrekonstruktion aufgrund paraverbaler Charakteristika gesprochener Sprache (vgl. zusammenfassend etwa Imhof 2003, S. 30–33). Dazu gehören zahlreiche Merkmale wie etwa Sprechtempo bzw. -dehnung, Pausen, Stimmklang und Tonhöhe, Lautstärke und Artikulation sowie Betonung (Akzent) auf Silben-, Wort-, (Teil-)

Satz- und Textebene, die jeweils für sich genommen, aber auch in Kombination die Bedeutung des Gesprochenen festlegen und z. T. in Nuancen differenzieren (vgl. auch Bose et al. 2013, S. 39). Will man die Fähigkeit überprüfen, eine so verstandene *eigentliche* Bedeutung zu ermitteln, so setzt das eine gewisse Eindeutigkeit der Zuordnung voraus. Dass der Zusammenhang zwischen paraverbalem Mittel und emotionalem Gehalt einer Äußerung nicht beliebig ist, zeigt sich in den Arbeiten von Baldur Neuber (2010): Seine Probanden hatten das Pseudowort *katakamala* jeweils freundlich, sachlich, eindringlich oder ärgerlich zu sprechen. Dabei produzierten sie melodische Konturen mit charakteristischen Merkmalen, die als Prototypen des Ausdruckes zumindest einiger grundlegender Emotionen gelten können, und die überindividuell mit großer Trefferwahrscheinlichkeit reidentifiziert wurden.

Wenn Texte eigens für die Testentwicklung als Hörtexte produziert werden, ist dies bereits bei der akustischen Interpretation des Textes zu berücksichtigen, denn ob eine Textstelle als bedeutsam in Erinnerung bleibt, hängt auch davon ab, wie der Sprecher bzw. die Sprecherin sie gestaltet: Derselbe Inhalt auf der verbalen Ebene kann je nach stimmlicher Gestaltung sehr Unterschiedliches bedeuten.

Dieser Aspekt der Zuhörkompetenz ist bislang zwar theoretisch modelliert und auch in einigen Studien zur Grundlagenforschung empirisch fundiert worden (für einen Überblick vgl. Behrens 2010, S. 44f.), konnte bislang aber noch nicht in Testverfahren integriert werden. Das führte in der Vergangenheit zu testbasierten Modellierungen, die aus zuhördidaktischer Sicht als unterkomplex angesehen werden müssen. Aus Gründen der Testkonstruktion konnte die Fähigkeit, aufgrund paraverbaler Merkmale zwischen verschiedenen Bedeutungen zu unterscheiden, bislang nicht in die Beschreibung der vorliegenden Kompetenzstufenmodelle aufgenommen werden.

Diese Lücke soll mithilfe der Projektergebnisse geschlossen werden. Ganz bewusst wurden Testaufgaben hergestellt, die neben den bislang üblichen Formaten auch solche Items enthalten, in denen die Testpersonen paraverbale Merkmale im angedeuteten Sinne verstehen müssen. Dabei wurden verschiedene *paraverbale* Itemtypen erprobt, die hier nur knapp charakterisiert werden sollen (für detailliertere Information vgl. die technische Dokumentation).

Typ 1: Eine Äußerung wird auf vier unterschiedliche Weisen gesprochen. Es ist die Option auszuwählen, die einer im Itemstamm vorgegebenen Sprecherintention (hier: leicht genervt) entspricht:

Die Königin ist leicht genervt, weil der König so laut ist. Wo hörst du das am besten?

☐ A) Weshalb weckst du mich mitten in der Nacht?
☐ B) Weshalb weckst du mich mitten in der Nacht?

☐ C) Weshalb weckst du mich mitten in der Nacht?
☐ D) Weshalb weckst du mich mitten in der Nacht?

Es wurde darauf geachtet, dass die Äußerung zwar zum Kontext der Geschichte passt, aber im Text nicht identisch enthalten ist. Auf diese Weise kann ausgeschlossen werden, dass Testpersonen nur ein bestimmtes klangliches Gebilde wiedererkennen, indem sie bemerken, dass sie die richtige Lösung schon einmal gehört haben.

Im Falle dieses Itemtyps ist es unabdingbar, dass der Hörtext selbst produziert wurde, denn der Interpret des Textes muss auch die vier Antwortoptionen mit den verschiedenen emotionalen Färbungen einsprechen. Erfahrungsgemäß ist hier eine enge Zusammenarbeit mit den Sprecherinnen und Sprechern erforderlich, die ggf. mehrfache Rückkopplungen ermöglicht. (Bewährt hat sich, in einer gemeinsamen Sitzung aus mehreren aufgezeichneten Versuchen die besten auszuwählen und diese im professionellen Tonstudio endgültig zu produzieren.) Dabei ist nicht nur zu beachten, dass die richtige Option eindeutig zuzuordnen ist, sondern auch, dass keiner der Distraktoren zu leicht ausgeschlossen werden kann (s. o.). Dies wäre beispielsweise der Fall, wenn drei negative und eine positive Stimmung zu hören wären *und* letztere die richtige Lösung darstellte. Gleichzeitig dürfen aber auch nicht mehrere Optionen allzu ähnlich klingen, also etwa Varianten oder Abstufungen der weiter oben vorgestellten Prototypen im Sinne Neubers sein, da eine eindeutige Korrespondenz zwischen paraverbalem Signal und intendierter Bedeutung nicht existiert. Der beträchtliche Produktionsaufwand für diesen Itemtypus ist dadurch gerechtfertigt, dass hier keine Möglichkeit besteht, die richtige Lösung allein auf Grundlage der verbalen Information zu finden.

Typ 2: Eine geeignete Äußerung aus dem Originaltext wird in den Itemstamm hineingeschnitten und es werden vier mögliche Bedeutungen dieser Äußerung in Form verschiedener Anschlussätze angeboten. Dieser Typus lässt sich auch auf Basis eines bereits fertigen Hörtextes herstellen. Er eignet sich also auch beispielsweise für professionell produzierte Hörspiele, Radiosendungen etc. Wiederum sind auf der verbalen Ebene alle vier Optionen möglich, erst auf der paraverbalen Ebene entscheidet sich die Passung mit der intendierten Stimmung. Die Regieanweisungen für den Sprecher stehen kursiv in Klammern hinter der jeweiligen Option; sie erscheinen nicht im Testheft.

Der König sagt zum Leibarzt: „Der Prinz, mein geliebter Sohn, ist verschwunden." Was könnte er am ehesten anfügen?

☐ A) „Machen wir uns auf die Suche." *(fröhliche Aufforderung)*
☐ B) „Wenn ich ihn finde, hört er aber was." *(schimpfend)*

☐ C) „Er wird wohl bald wieder auftauchen." *(zufrieden, im Sinne von „halb so wild")*
☐ D) „Ich mache mir große Sorgen um ihn." *(besorgt wie im Itemstamm)*

Bei diesem Itemtyp ist es möglich, die Formulierung des Itemstamms (außer der übernommenen Äußerung) und der Antwortoptionen von einem anderen Sprecher aufnehmen zu lassen.

Typ 3: Die einer Äußerung zuzuordnende Intention auf metasprachlicher Ebene wird expliziert. Auch dieser Itemtyp eignet sich für bereits fertiggestellte Hörtexte. Es ist allerdings zu beachten, dass die Formulierung der fraglichen Intention(en) in für Kinder verständlicher Weise möglich ist, was sich häufig als schwierig erweist. Ein Beispiel für diesen Typus ist das folgende:

Der König sagt: „Ich wünsche, dass das Schloss von oben bis unten durchsucht wird!" Wie meint er das?

☐ A) Das ist ein Befehl.
☐ B) Das ist eine Bitte.
☐ C) Das ist eine Entschuldigung.
☐ D) Das ist eine Drohung.

Hier könnten theoretisch Wörter wie *Befehl* oder *Drohung* zu einem Wortschatzproblem führen – es wäre also denkbar, dass ein Kind zwar die Sprecherintention richtig erfasst, aber dennoch die falsche Option auswählt, weil es den treffenden Ausdruck nicht zuordnen kann. Diese Möglichkeit lässt sich nicht vollkommen ausschließen. Es wurde aber eine Variante erprobt, in der die gefragte Intention umschrieben wird wie in diesem Beispiel:

Die Köchin sagt: „Vielleicht ist er bloß weggelaufen. Kinder laufen manchmal von zu Hause weg." Weshalb sagt sie das so?

☐ A) Sie weiß, dass der Prinz weggelaufen ist.
☐ B) Sie will sich wichtig machen.
☐ C) Sie möchte die anderen beruhigen.
☐ D) Sie möchte endlich wieder ins Bett gehen.

Empirisch zeigt sich allerdings zwischen diesen beiden Typen – zumindest in dieser Aufgabe – kein Unterschied: Beide Items sind der Kompetenzstufe I zuzuordnen; das erste wird mit 89% noch etwas häufiger gelöst als das zweite (76%). Eine systematische Analyse steht aber auch hier noch aus.

9 Zusammenfassung

Im Projekt *stɪm·mig* wurden Testaufgaben zur Messung der Zuhörfähigkeit von Grundschulkindern entwickelt und im Rahmen einer Großstudie evaluiert. Im Prozess der Entwicklung zeigte sich, dass zu allen Zeitpunkten mehrere Fachperspektiven in Betracht gezogen werden mussten, um zu (theoretisch) gelungenen Aufgabenformulierungen zu kommen. Diese Perspektiven wurden im Artikel anhand eines konkreten Aufgabenbeispiels dargelegt und sind unten in einer Tabelle (Tab. 1) stichpunktartig zusammengeführt. Ein wesentliches Charakteristikum solcher interdisziplinären Projekte ist, dass die Perspektiven nicht mehr einzelnen Personen(-gruppen) strikt zugeordnet werden können. Vielmehr ergibt sich im gemeinsamen Arbeitsprozess erkennbar eine gegenseitige Qualifikation aller Beteiligten. Dass solche Projekte vermutlich in Zukunft den Normalfall empirischer Forschungs- und Entwicklungsvorhaben im Feld der Deutschdidaktik darstellen, kann mithin nur als erfreulicher Umstand betrachtet werden, der zu einer weiteren Professionalisierung der Disziplin beitragen wird.

Tab. 1: Fachperspektiven im Projekt anhand eines konkreten Aufgabenbeispiels

	literaturwissenschaftliche und sprachdidaktische Perspektive	**psychometrische Perspektive**	**schauspielerische/ sprechwissenschaftliche Perspektive**	**administrativ-technische Perspektive**
Textauswahl	Ist der Text inhaltlich und sprachlich für die Zielgruppe geeignet?	Ist der Text für alle Probanden gleichermaßen (un-) vertraut?	Wie kann der Text insgesamt angemessen inszeniert werden (Grundstimmung, Figurengestaltung)?	Bestehen urheberrechtliche Bedenken?
Textausschnitt	Funktioniert der gewählte Ausschnitt im Sinne eines in sich abgeschlossenen Textes?	Ist das Verhältnis von Textlänge und Zahl der Testitems angemessen (Testökonomie)?	Wie viele und welche Sprecher werden benötigt (Geschlecht, Alter…)?	Eigenproduktion oder externe Auftragsvergabe? Vertragsgestaltung?
Testfragen	Welche Informationen sind relevant für ein angemessenes Textverständnis? Wie sind diese im Gedächtnis repräsentiert?	Zielen die Fragen auf unterschiedliche Ebenen des getesteten Konstrukts ab (verschiedene Anforderungsbereiche)?	Wie un-/ähnlich müssen bzw. dürfen Antwortoptionen klingen? Können Emotionen und Bedeutungen stimmlich erkennbar markiert werden?	Können alle Fragen vom gleichen Sprecher eingespielt werden? Sind Schneidearbeiten im Tonstudio erforderlich?

	literaturwissenschaftliche und sprachdidaktische Perspektive	**psychometrische Perspektive**	**schauspielerische/ sprechwissenschaftliche Perspektive**	**administrativ-technische Perspektive**
Item-Konstruktion	Welche Itemvarianten (z. B. Modalität: nur hören vs. hören + mitlesen) sind zur Hypothesenprüfung nötig?	Welche Itemformate sollen gewählt werden? Formulierung eindeutiger Auswertungsanweisungen	-	Wie aufwändig ist die Auswertung? Sind (ggf. mehrere) Kodiererschulungen erforderlich?
Testdesign	-	„Rotierte" Anordnung der Aufgaben/-blöcke in Testheften (zur Vermeidung von Reihenfolgeeffekten)	-	Stehen ausreichend Testpersonen zur Verfügung? Ist die Stichprobe repräsentativ?

Literatur

Baumert, Jürgen et al. (Hrsg.) (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Behrens, Ulrike (2010): Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung. In: Bernius, Volker/Imhof, Margarete (Hrsg.): Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis. Göttingen: Vandenhoeck & Ruprecht, S. 31–50.

Behrens, Ulrike/Böhme, Katrin/Krelle, Michael (2009): Zuhören. Operationalisierung und fachdidaktische Implikationen. In: Granzer, Dietlinde/Köller, Olaf/Bremerich-Vos, Albert (Hrsg.): Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule. Weinheim: Beltz, S. 357–376.

Böhme, Katrin (2011): Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen. Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich. http://edoc.hu-berlin.de/dissertationen/boehme-katrin-2011-12-16/PDF/boehme.pdf. Abgerufen am 05.10.2015.

Böhme, Katrin/Robitzsch, Alexander/Busè, Anne-Kathrin (2010): Zur Abgrenzung des Hörverstehens. In: Bernius, Volker/Imhof, Margarete (Hrsg.): Zuhörkompetenz in Unterricht und Schule. Göttingen: Vandenhoeck und Ruprecht, S. 81–104.

Bose, Ines et al. (Hrsg.) (2013): Einführung in die Sprechwissenschaft. Phonetik, Rhetorik, Sprechkunst. Tübingen: Narr Studienbücher.

Buck, Gary (2001): Assessing listening. Cambridge: Cambridge University Press.

Fiehler, Reinhard (2014): „Von der Mündlichkeit zur Multimodalität … und darüber hinaus." In: Grundler, Elke/Spiegel, Carmen (Hrsg.): Konzeptionen des Mündlichen. Wissenschaftliche Perspektiven und didaktische Konsequenzen. Bern: hep Verlag, S. 13–31.

Fiehler, Reinhard et al. (2004): Eigenschaften gesprochener Sprache. Theoretische und empirische Untersuchungen zur Spezifik mündlicher Kommunikation. Tübingen: Narr.

Grotjahn, Rüdiger (2000): Testen der Fertigkeit Hörverstehen. Leistungsmessung und Leistungsbeurteilung. http://www.unileipzig.de/herder/temp/lehrende/tschirner/testen/hoeren.pdf. Abgerufen am 05.10.2014.

Imhof, Margarete (2003): Zuhören. Psychologische Aspekte auditiver Informationsverarbeitung. Göttingen: Vandenhoeck & Ruprecht.

IQB (2013): Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich „Sprechen und Zuhören" – Primarbereich – Beschluss der Kultusministerkonferenz (KMK) vom 04.03.2010. Auf Grundlage des Ländervergleichs 2011 überarbeiteter Entwurf in der Version vom 13. Februar 2013. https://www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Deutsch_Z.pdf. Abgerufen am 05.10.2014.

Moosbrugger, Helfried/Kelava, Augustin (2012): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer Verlag.

Müller, Karla (2012): Hörtexte im Deutschunterricht. Poetische Texte hören und sprechen. Seelze: Klett Kallmeyer.

Neuber, Baldur (2002): Prosodische Formen in Funktion. Leistungen der Suprasegmentalia für das Verstehen, Behalten und die Bedeutungs(re)konstruktion. Frankfurt a. M.: Lang.

Nold, Günter/Rossa, Henning (2007): Hörverstehen. In: Beck, Bärbel/Klieme, Eckhard (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim: Beltz, S. 178–196.

Quasthoff, Uta (1980): Erzählen in Gesprächen. Tübingen: Narr.

Schnotz, Wolfgang (1996): Lesen als Textverarbeitung. In: Günther, Hartmut/Ludwig, Otto (Hrsg.): Schrift und Schriftlichkeit. Writing and its Use. Ein interdisziplinäres Handbuch internationaler Forschung. Berlin/New York: de Gruyter, S. 972–982.

Zingg Stamm, Claudia/Käser-Leisibach, Ursula/Bertschin, Felix (2014): Ohrwärts. Zuhören und literarisches Hörverstehen. Kompetenzerhebung mit Förderangeboten für 9- bis 10-Jährige. Solothurn: Lehrmittelverlag Solothurn.