MODUL 2

TESTENTWICKLUNG

1. **CHECKLISTE FÜR DIE BEURTEILUNG VON TESTS**
2. **TESTBESCHREIBUNG (ENGL. SPECIFICATIONS)**
3. **PRODUKTIONSPROZESS**
   a) Auftragsvergabe
      i) Hinweise zur Auswahl der Texte
      ii) Hinweise zur äußeren Form
      iii) Detaillierte Hinweise zu jeder Aufgabe
   b) Begutachtung und Redaktion
   c) Erprobung
   d) Statistische Analysen der Items
      i) Schwierigkeit der Items
      ii) Trennschärfe der Items
      iii) Angaben zu den Distraktoren
   e) Kalibrierung der Items
      i) Erstellen einer Schwierigkeitsskala
      ii) Verankerung
      iii) Interpretation der Scores
   f) Itembank
      i) Itembank und Testzusammenstellung
      ii) Beispiel für eine Itembank-Software
      iii) Testzusammenstellung ohne Itembank
4. **ÜBUNGEN**

ANHANG A – Abgelehnte Texte

ANHANG B – Iteman-Beispiele (MicroCAT) für Übung 4

ANHANG C – Literaturempfehlungen

ANHANG D – Ein Beispiel für die Testentwicklung,
Entwicklung eines Tests für die Fertigkeit Sprechen

TESTENTWICKLUNG

Alle Tests, unabhängig von ihrem Ziel oder ihrem Niveau, müssen bestimmten Kriterien genügen. Deshalb ist es sinnvoll, zuerst diese Kriterien genauer zu betrachten, bevor untersucht wird, wie Tests erstellt werden. Die folgenden Fragen liegen dem gesamten Prozess der Testentwicklung zugrunde.

1. Checkliste für die Beurteilung von Tests

Checkliste für die Beurteilung von Tests

Validität des Tests

Was ist das Ziel des Tests?
Stimmt der Inhalt des Tests mit den Zielen überein, die für den Einsatz des Tests genannt werden – ist er valide?

Schwierigkeit des Tests

Welche Merkmale haben die Testteilnehmer?
Ist der Test angemessen im Hinblick auf die gesamten Fähigkeiten der Teilnehmer?
Wurde der Test mit einer Probandengruppe erprobt, die repräsentativ für die Population der Testteilnehmer ist?

Reliabilität des Tests

Sind die Ergebnisse im Test zuverlässig genug, um präzise Entscheidungen zu treffen?
Kann der Grad an Reliabilität nachgewiesen werden? Wie messgenau ist der Test?

Anwendbarkeit des Tests

Inwieweit sind Format und Merkmale des Tests im Hinblick auf das Umfeld seines Einsatzes angemessen?
Wie vertraut sind die Testteilnehmer mit dem Testformat?
Können Format und Merkmale beim Echteinsatz des Tests angemessen und fair umgesetzt werden?

Relevanz des Tests

Wie relevant ist der Test für die Zielgruppe und/oder den Bereich, den die Items erfassen sollen?
Inwieweit stimmt die avisierte Zielgruppe/der Inhalt mit der Auswahl überein, die während der Entwicklungsphase getroffen wurde?

Wiederholbarkeit des Tests

Wie einfach ist es, äquivalente oder gleichgesetzte Versionen des Tests zu erstellen?

Interpretation der Testergebnisse

Wie einfach ist es, die Testergebnisse zu ermitteln, auszuweisen und zu interpretieren?
Ist dazu ein erheblicher Trainingsaufwand nötig?

Testökonomie
Was kostet es, den Test bereitzustellen, ihn durchzuführen und die Ergebnisse zu bewerten?

Verfügbarkeit des Tests
Ist der Test leicht zugänglich?

Angemessenheit des Tests
Ist der Test angemessen im Hinblick auf die durchführende Institution oder die Gesellschaft?
Halten die Lehrer, die Eltern und die Testdurchführenden den Test für angemessen?

(Nach: Grant Henning, *A Guide to Language Testing,* Newbury House, 1987)

Bei der Beschreibung des Entwicklungsprozesses von Tests ist es wichtig, darauf hinzuweisen, dass diejenigen, die die Items für den Test schreiben, in unterschiedlichem Maß an diesem Prozess beteiligt sind. Dies kann sich auf einen Arbeitsvertrag als freier Mitarbeiter für das Schreiben von Items beschränken, möglicherweise eine Beteiligung an der Revision und Fertigstellung der Endversion eines Items mit einschließen oder sogar die Verantwortung für alle Phasen der Testentwicklung beinhalten.

Es ist deshalb dringend erforderlich, nicht nur den Erstellungsprozess von Prüfungsaufgaben zu bedenken, sondern auch den gesamten Entwicklungsprozess dieser Prüfung. Beides sind zyklische Prozesse, in die das erworbene Wissen und die gewonnene Erfahrung eingebunden werden. Desgleichen fließen nach dem Einsatz der Prüfung Ergebnisse der fortwährenden Evaluierung und Durchführung mit in diese Prozesse ein.

Ein Modell für den gesamten Entwicklungsprozess einer neuen Prüfung findet sich in Abbildung 1 auf der folgenden Seite. Dieses Diagramm zeigt exemplarisch, welche Phasen durchlaufen werden müssen, beginnend mit der Erkenntnis, dass eine neue Prüfung gebraucht wird, bis zur tatsächlichen Einsatz- und Umsetzungsphase, in die die Testautoren am stärksten involviert sind. Ein ausführliches Beispiel für den gesamten Entwicklungsprozess findet sich in Anhang D.

Wie bereits gesagt, ist die Testentwicklung ein zyklischer Prozess. Sobald die Notwendigkeit für einen neuen Test erkannt wird, führt dies zur Planungsphase, in der Daten über die Bedürfnisse der Kandidaten gesammelt werden. Dies kann mithilfe von Fragebögen geschehen und u. a. in den Institutionen stattfinden, die den Test später vermutlich einsetzen werden. Dabei sollte ein klares Bild entstehen, wer die zukünftigen Kandidaten voraussichtlich sind. Während der darauffolgenden Konzeptionsphase wird ein erster Entwurf der Testbeschreibung erstellt (d. h. der detaillierten Beschreibung der Prüfungsziele, des Testformats, der Itemtypen, der Texte und der Bewertungsverfahren), die für diese Kandidaten geeignet ist. Das Design des Tests und alle Aspekte seines Inhalts werden diskutiert, zusammen mit den Gesichtspunkten und Hemmnissen, die sich darauf auswirken. Entscheidungen werden gefällt hinsichtlich der Länge der einzelnen Testteile, der ausgewählten Itemtypen und der zur Verfügung stehenden Themen. Beispielmaterialien werden erstellt.

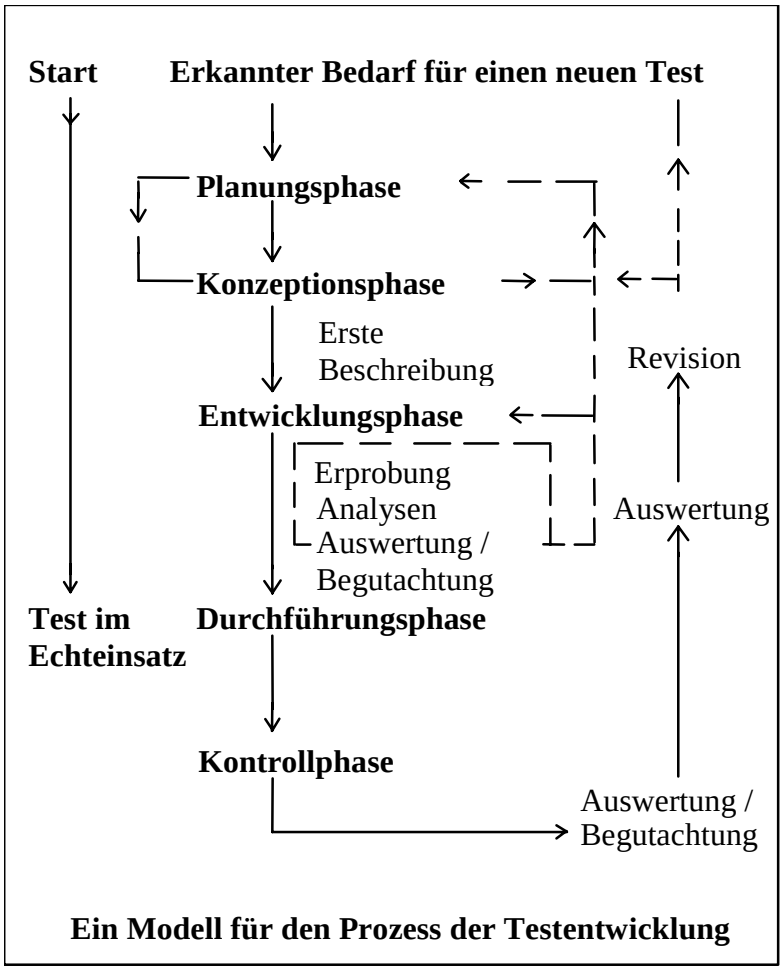

Abbildung 1: Modell für die Testentwicklung (Saville und Milanovic)

Während der Entwicklungsphase werden die entwickelten Materialien erprobt. Dies bedeutet, dass Lerner mit entsprechendem sprachlichen Niveau und auch anderen Merkmalen zukünftiger Kandidaten (z. B. das Alter) den Test unter Prüfungsbedingungen ablegen. Lösungsschlüssel und Bewertungsanleitungen werden erstellt. Mithilfe von Fragebögen und Stellungnahmen der Lerner und ihrer Lehrer werden Informationen zum Test gesammelt, Video- und Tonaufnahmen und Beobachtungen werden durchgeführt und die Ergebnisse analysiert. Daran anschließend können Entscheidungen darüber gefällt werden, ob die Materialien das richtige Schwierigkeitsniveau haben und sich auch sonst für den Echteinsatz in einem Test eignen. Zu diesem Zeitpunkt ist es noch möglich, erhebliche Änderungen an der Testbeschreibung, den gewählten Aufgabentypen sowie an allen anderen Aspekten des Tests vorzunehmen.

Wenn diese ersten Phasen abgeschlossen sind, wird die Testbeschreibung in ihre endgültige Form gebracht, die Aufgaben geschrieben und der Test zusammengestellt. Durchführungs- und Bewertungsbestimmungen werden festgelegt. Dies ist die Einsatz- oder Durchführungsphase, in der der Test für Kandidaten verfügbar wird. Diese Phase wird in Abbildung 3 ausführlicher dargestellt.

Die Ergebnisse der Testdurchführung werden überwacht und Rückmeldungen von den Kandidaten und Lehrern der Institutionen, die den Test durchführen, eingeholt. Solche Daten werden verwendet, um die Qualität des Tests zu beurteilen und die Notwendigkeit für eine Revision zu ermitteln. Verschiedene Aspekte des Kandidaten- und Prüferverhaltens können untersucht werden, um festzustellen, welche Änderungen am Test oder seinen Durchführungsmodalitäten vorgenommen werden müssen. Ab einem bestimmten Zeitpunkt wird eine Revision ohnehin erforderlich, um neue Entwicklungen in Testtheorie und Testpraxis aufzugreifen. Damit kehrt der Test zum Ausgangspunkt des Kreislaufs zurück, denn eine umfangreiche Testrevision bedeutet, wieder bei der Planungsphase zu beginnen.

2. Testbeschreibung (engl. *specifications*)

Wenn die Testbeschreibung für einen neuen (oder revidierten) Test entwickelt wird, geschieht dies immer mit dem Ziel, eine **valide** und **reliable** Prüfung zu erstellen, die eine **positive Auswirkung** hat und **praktikabel** ist. Mit anderen Worten sollte der Test ein angemessenes Leistungsmessungsverfahren sein, das misst, was gemessen werden soll; die Ergebnisse sollten möglichst frei von Messfehlern sein; der Test sollte möglichst positive Auswirkungen auf das Individuum und den Unterricht haben und die damit verbundenen Anforderungen an Mittel und Möglichkeiten der Testentwickler sollten zu den zur Verfügung stehenden Mitteln passen. Während der Planungsphase sollten diese Faktoren immer im Auge behalten und eine Balance zwischen ihnen hergestellt werden.

Die erste Phase in der Planung bezieht sich auf die Situationsanalyse, d. h., man muss den Bedarf einer neuen Prüfung im Kontext der verschiedenen Einflüsse analysieren, die auf sie einwirken und die ihre endgültige Form mitbestimmen. Dabei sollten die wichtigsten Aspekte und Beschränkungen festgehalten werden, die für das Projekt relevant sind. Dazu gehören die Frage, wie ein Test beschaffen sein muss, um seinem Ziel und seinem Zweck zu genügen, und die Beschränkungen, denen ein Test durch die Umstände seiner Durchführung unterworfen ist.

Die Überlegungen sind somit sowohl fachlicher als auch praktischer Natur.

Die **fachlichen Überlegungen** betreffen die im Test zu prüfenden Inhalte. Sie umfassen:

- die Realsituationen, in denen die Kandidaten die Sprache gebrauchen müssen
- das Niveau der Sprachbeherrschung, das für diese Situationen erforderlich ist
- die realen kommunikativen Aktivitäten, die der Test abbilden muss
- die Informationen, die dem Benutzer des Tests gegeben werden müssen

Die **praktischen Überlegungen** betreffen die Beschränkungen, denen die Leistungsmessung unterworfen ist, wie z. B.:

- das zur Verfügung stehende Personal und die vorhandenen Räume
- die Teilnehmerzahl und die Länge des Tests
- die Verfügbarkeit angemessen geschulter Prüfer
- die Aufgabenformate, deren Verwendung wünschenswert erscheint
- das Verfahren, mit dem den Kandidaten ihre Ergebnisse mitgeteilt werden
- die Verfahren der Qualitätskontrolle, die vorgesehen sind

Die Beschränkungen betreffen:

- die Akzeptanz der Prüfung durch alle von ihr Betroffenen, d. h. Kandidaten, Eltern, Lehrer, Schulleiter usw.
- die Passgenauigkeit der Prüfung hinsichtlich Curriculum und Unterricht
- den erforderlichen Schwierigkeitsgrad
- die Erwartungen, wie eine solche Prüfung aussehen sollte, inklusive möglicher Vergleiche mit anderen ähnlichen Produkten auf dem Markt
- die Verfügbarkeit von Mitteln für die Entwicklung und Durchführung sowie für die Bekanntgabe der Ergebnisse

Wie Abbildung 1 zeigt, wird nach Fertigstellung der Testbeschreibung ein erster Entwurf des Tests erstellt und Beispielmaterial entwickelt. Dieses Material wird dann mit Lernern erprobt, die das adäquate sprachliche Niveau haben, und die Ergebnisse werden analysiert. Als Ergebnis der Erprobung werden möglicherweise einzelne Aufgabentypen oder bestimmte Materialien verworfen und die Länge der Testteile oder Aspekte der Durchführung evtl. modifiziert. Die Testbeschreibung muss evtl. mehrere Revisionen durchlaufen, bevor sie ihre endgültige Form erhält.

Es ist möglich, dass ein und dieselbe Person sowohl für die Entwicklung der Testbeschreibung als auch später für das Erstellen der Testaufgaben des einsatzbereiten Tests verantwortlich ist. Aber auch zuvor nicht beteiligte Personen müssen der Testbeschreibung detaillierte Informationen über die Prüfung entnehmen können, sei es, weil sie Items für den Test schreiben sollen, sei es, weil sie diese Informationen brauchen, um zu entscheiden, ob sie ihre Lerner anmelden oder nicht. Ein Testautor, der noch keine Items für eine Prüfung entwickelt hat und der an den Entwicklungsphasen nicht beteiligt war, braucht klar definierte Vorgaben, an die er sich halten kann; die Testbeschreibung muss diesen Zweck weitestgehend erfüllen.

Die Testbeschreibung in ihrer endgültigen Fassung sollte detaillierte Informationen über jeden Testteil oder Subtest geben. Sie sollte die folgenden Aspekte behandeln:

- wie lange der Test dauert
- in wie viele Teile er untergliedert ist
- aus wie vielen Items jeder Teil besteht
- die in jedem Teil verwendeten Aufgabentypen
- der Fokus eines jeden Items, wie z. B. detailliertes Textverstehen
- was getestet wird, z. B. Anwendung grammatischer Regeln
- die Art der Texte, die als Input verwendet wird
- die Gesamtlänge der verwendeten Texte und die Länge der einzelnen Texte
- die Textquellen
- Angaben über die Themenbereiche, die für angemessen gehalten werden
- das Format und der Umfang der Aufgaben
- der Punktwert, der für jedes Item vergeben wird, und der maximal zu erreichende Punktwert
- genaue Angaben zur Gewichtung
- bei einer Bewertung durch Beurteiler/Prüfer: Anleitung zu Bewertungsvorgaben und zur Koordination der Beurteiler-/Prüferteams
- genaue Angaben über die Kriterien bei der Bewertung des Schreibens und des Sprechens
- wie viele Beurteiler oder Prüfer beteiligt sind, z. B., ob routinemäßig eine zweite Beurteilung stattfindet
- Art der Vorgaben, die in Tests zum Sprechen verwendet werden
- genaue Angaben zum Bewertungsverfahren und zur Bekanntgabe der Ergebnisse
- wo und wann der Test abgelegt werden kann
- Verfügbarkeit von Modelltests
- geschätzte Stundenzahl, um sich auf die Prüfung vorzubereiten

All diese Informationen helfen den Testautoren, sich ein klares Bild davon zu machen, wie die Materialien beschaffen sein müssen, an deren Produktion sie beteiligt sind.

Manchmal ist es hilfreich, die wesentlichen Informationen über einen Test einer kurzen Zusammenfassung entnehmen zu können. Ein Beispiel für eine Zusammenfassung in Form eines Rasters findet sich unten. Die Prüfung, auf die sich das Raster bezieht, besteht aus fünf Subtests. Jeder Subtest wird zusammengefasst, indem beschrieben wird, worauf der Fokus liegt, wie der gegebene Input beschaffen ist und wie die erwartete Antwort aussieht.

Teil 1 Leseverstehen	Testfokus	Input	Format
	Erkennen der Angemessenheit von Grammatik und Lexik. Hauptinhalte eines Textes und die Funktion seiner Aussagen verstehen. Die wichtigsten Punkte trotz gelegentlicher Wortschatzlücken erfassen. Einem Text spezifische Inhalte entnehmen. Meinungen und Einstellungen erkennen, wenn sie klar geäußert werden. Detailverstehen.	Abschnitt A: Einzelsätze Abschnitt B: Drei oder vier Texte unterschiedlicher Texttypen: narrativ, beschreibend, darstellend, diskursiv, informativ usw. Mögliche Quellen sind: literarische Texte, Sachtexte, Zeitungen, Zeitschriften, Werbung, Informations-broschüren usw.	Abschnitt A: 25 einzelne Multiple-Choice-Items mit je vier Optionen. Abschnitt B: 15 Multiple-Choice-Items mit je vier Optionen, die sich auf die drei oder vier Texte beziehen.
Teil 2 Aufsatz	Testfokus	Input	Format
	Natürlich und angemessen schreiben, um auf eine Reihe thematischer oder situationsbezogener Stimuli zu antworten.	Vier kurze situationsbezogene Vorgaben oder Aufgaben über eine Reihe alltäglicher Themen. Je eine Aufgabe zur vorgeschriebenen Lektüre (drei Bücher).	Zwei von fünf Aufgaben auswählen und bearbeiten; erforderter Umfang der Lösung: je 120-180 Wörter. Geforderte Textsorten beinhalten Briefe, beschreibende/erzählende Texte, ausführlichere Abhandlungen und geschriebene Reden.
Teil 3 Gebrauch des Englischen	Testfokus	Input	Format
	Gebrauch des Englischen auf der Wort- und Satzebene inklusive Verwendung korrekter Strukturen: korrekte und angemessene Wörter und Sätze; Vielfalt an Formen, um ähnliche Bedeutung auszudrücken; Wortableitungen. Informationen korrekt und angemessen schriftlich zusammenfassen.	Aufgaben auf der Basis kurzer Texte oder Einzelsätze. Auch visueller Input (Karten, Diagramme usw.) in gelenkten Aufgaben zum Schreiben.	Lückentest, Umformungsaufgaben, Aufgaben zur Wort- und Satzbildung. Gelenkte Schreibaufgabe.

Teil 4 Hörverstehen	Testfokus	Input	Format
	Verstehen der Hauptinformationen in einem gehörten Text. Seine Funktion und Aussage verstehen. Die wichtigen Inhalte trotz Wortschatzlücken verstehen. Dem gehörten Text spezifische Informationen entnehmen. Tonfall und Einstellung erkennen, wenn sie klar geäußert werden. Details in einem gehörten Text verstehen.	Drei oder vier authentische oder semi-authentische Aufnahmen. Mögliche Textquellen: Nachrichten, Programme, Meldungen, Unterhaltungen, öffentliche Reden, Ansagen usw.	Drei oder vier Aufgaben, insgesamt ungefähr 30 Items. Mögliche Aufgabentypen: Multiple-Choice, Lücken ergänzen, Notizen machen, Richtig/Falsch, Ja/Nein etc.
Teil 5 Sprechen	Testfokus	Input	Format
	Sich an einem Gespräch auf Englisch über eine Reihe von Themen beteiligen, die von alltäglichen bis zu abstrakten Themen reichen können. Beim Sprechen angemessene Flüssigkeit, Interaktion, Aussprache auf Wort- und Satzebene, Korrektheit und Angemessenheit des Wortschatzes nachweisen.	Mögliche Vorgaben: Fotos, kurze Texte, visuelle Stimuli. Die Vorlagen können sich auf eine Liste empfohlener Lektüre beziehen.	Ein Gespräch zwischen Prüfer und Kandidat(en) über ein Thema. Besteht aus drei Abschnitten: 1. Über ein Foto oder mehrere Fotos sprechen 2. Über einen kurzen Text sprechen 3. Eine kommunikative Tätigkeit ausführen. Das Gespräch kann als Einzel-, Paar- oder Gruppenprüfung (drei Kandidaten) durchgeführt werden.

Abbildung 2: Informationen über eine Prüfung in Form eines Rasters

3. Produktionsprozess

Die Testbeschreibung bestimmt, welches Material für den Test erstellt werden muss. Der eigentliche Produktionsprozess besteht oft aus fünf Schritten:

- Vergabe des Arbeitsauftrages
- Bearbeitung
- Erprobung
- statistische Analysen und Aufnahme in eine Itembank
- Zusammenstellen des einsatzbereiten Tests

Dieser Prozess wird in Abbildung 3 auf der nächsten Seite dargestellt. Bei der Betrachtung der Grafik muss berücksichtigt werden, dass ein Modell für objektiv auswertbare Testaufgaben etwas von dem Modell abweicht, das für die notwendigerweise subjektiver bewerteten Testaufgaben zum Sprechen oder zum „freien“ Schreiben verwendet wird, da Letztere nicht auf die gleiche Art und Weise erprobt werden können.

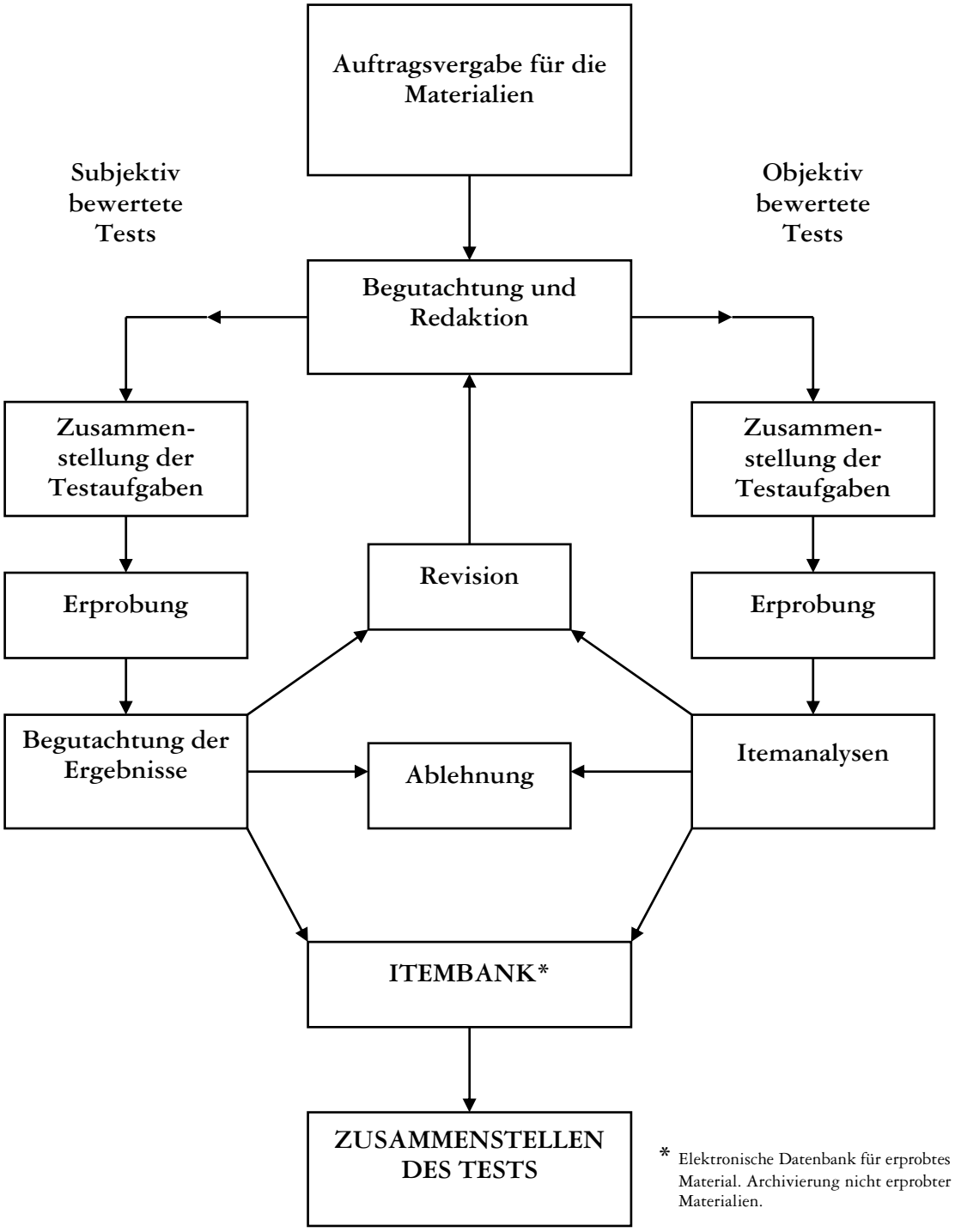

Abbildung 3: Die Produktionsphasen von Tests

In allen Phasen des Produktionsprozesses, unabhängig davon, wie viele Leute dran beteiligt sind, müssen zwei Regeln beachtet werden:

- Zeitmanagement: Dies bedeutet, dass realistische Projektpläne aufgestellt und Abgabetermine strikt eingehalten werden.
- Festhalten der Ergebnisse: Dies ist bei jedem Prozess unbedingt erforderlich, in dem es um Revisionen, Veränderungen und mehrmaliges Überarbeiten der Materialien geht.

a) Auftragsvergabe

Wie schon erwähnt, kann ein und dieselbe Person (nennen wir sie Testentwickler) für alle Phasen des Produktionsprozesses, auch für das Schreiben der Testaufgaben, verantwortlich sein. Es ist jedoch in dieser Phase üblich, dass der Testentwickler Arbeitsaufträge für Texte und/oder Items an verschiedene Personen vergibt, die entweder für die gleiche Organisation arbeiten oder externe freie Mitarbeiter mit Unterrichts- oder Prüfungserfahrung sind.

Manchmal ist ein Mitarbeiter einer Organisation, die Prüfungen erstellt, verantwortlich für die Vergabe von Arbeitsaufträgen, für die Bearbeitung der Entwürfe und dafür, die Items zu einer Prüfung zusammenzustellen, während sich andere um die Erprobungen, die statistischen Analysen und die Itembank kümmern. Die gleiche Person kann für alle Teile einer Prüfung verantwortlich sein oder im Falle einer Prüfung mit Subtests zum Lesen, Schreiben, Hören und Sprechen können verschiedene Personen für je einen Prüfungsteil zuständig sein.

Die Vergabe von Arbeitsaufträgen kann einem festgelegten Rhythmus folgen und z. B. zweimal im Jahr stattfinden oder sie kann dann erfolgen, wenn der Testentwickler es für notwendig erachtet, neues Material entwickeln zu lassen.

Testautoren können mit der Erstellung von Aufgaben für eine komplette Prüfung oder nur bestimmte Prüfungsteile beauftragt werden.

Das Ziel des Testentwicklers ist es, möglichst viel Testmaterial zu erhalten, das sich nach der Begutachtung für die Erprobung eignet und sich danach als tauglich erweist, in die endgültige Prüfung aufgenommen zu werden. Es ist deshalb seine Aufgabe, gute Testautoren zu beauftragen und ihnen möglichst klare und sinnvolle Anweisungen zu geben. Man findet freie Testautoren oft im Kreise derer, die Erfahrungen mit der Prüfung haben, entweder weil sie ihre Lerner darauf vorbereiten oder weil sie als Prüfer oder Beurteiler für diese Prüfung tätig sind. Ob der Testentwickler nun alleine oder mit Kollegen arbeitet oder freie Mitarbeiter beauftragt, in der Phase der Auftragsvergabe sollten die folgenden Punkte klar sein:

- **Präzise Hinweise zu den in Auftrag gegebenen Materialien**
  Werden Texte verlangt, muss klar sein, ob die Items zusammen mit dem Text geliefert werden müssen oder erst, nachdem der Text angenommen wurde. Der Testautor sollte bei jedem Item die Lösung mitliefern. Es kann verlangt werden, eine Tonaufnahme und ein Manuskript für die Hörtexte mitzuliefern. Bei Tests zum Sprechen sollte klargestellt werden, ob der Testautor die visuellen Vorgaben mitliefern muss oder ob Hinweise dazu reichen, wie die Vorgaben aussehen sollen.

- **Hinweise zur Form der eingereichten Materialien**
  Handgeschriebene Materialien werden vermutlich nicht ausreichen, wohl aber Fassungen auf Diskette und/oder auf Papier. Wenn ein kompletter Prüfungsteil erstellt werden soll, muss der Testautor wissen, ob die Items durchgehend nummeriert werden und die Teile nacheinander folgen oder ob jeder Teil oder jede Aufgabe auf ein neues Blatt Papier geschrieben werden soll. Es ist nützlich, wenn die Testautoren ihren Namen, das Datum und den Namen der Prüfung oder des Subtests auf jedes Blatt schreiben (all dies kann in der internen Testbeschreibung oder den Hinweisen für Testautoren behandelt werden).

- **Hinweise zum Abgabetermin der Materialien**
  Es hilft den externen Testautoren zu wissen, wie ihre Arbeit in den gesamten Produktionsprozess eingebunden ist, damit klar wird, warum Abgabetermine eingehalten werden müssen. Wenn der Arbeitsauftrag vergeben wird, ist es i. d. R. möglich, den Testautoren den Zeitraum der Begutachtungsphase mitzuteilen. Darüber hinaus kann eine Information bzw. Anfrage bzgl. einer Teilnahme an der Begutachtung erfolgen.

- **Hinweise auf das Honorar**
  Von Anfang an sollte klar sein, unter welchen (finanziellen und anderen) Bedingungen die Testautoren arbeiten. Es ist möglich, dass nur für angenommene Materialien gezahlt wird und für abgelehnte Materialien kein Honorar anfällt oder dass beim Einreichen der Materialien nur eine geringe Summe gezahlt wird, die später für alle angenommenen

Materialien entsprechend aufgestockt wird. Es ist auch möglich, den Testautoren eine Liste der Honorare für die verschiedenen Aufgabentypen zu geben oder einfach nur das Gesamthonorar anzugeben, das für einen kompletten Prüfungsteil oder eine komplette Prüfung gezahlt wird.

Wenn die Testautoren einen Arbeitsauftrag erhalten, bekommen sie vermutlich auch einige der folgenden Unterlagen:

- die Testbeschreibung (interne und öffentliche, siehe unten)
- Beispielmaterialien oder einen Modellsatz der Prüfung
- ein Handbuch (oder Anweisungen bzw. einen Leitfaden) für Testautoren, die sich auf die betreffende Prüfung oder den entsprechenden Prüfungsteil beziehen
- ein Formular, das sie unterschreiben müssen, um zu erklären, dass sie den Arbeitsauftrag annehmen
- ein Formular, in dem festgehalten wird, dass der Auftraggeber das Urheberrecht für die eingereichten Materialien besitzt
- eine Liste oder ein Wörterbuch, die/das den Umfang und das Niveau des verwendeten Wortschatzes und/oder der grammatischen Strukturen festlegt
- allgemeine Handreichungen mit Informationen über die Institution, die die Prüfung(en) produziert

Die **öffentliche Testbeschreibung** dient der Information des allgemeinen Publikums. Obwohl sie die Inhalte der Prüfung sehr ausführlich darstellt, enthält sie keine detaillierten Hinweise zum Produktionsprozess oder zu besonderen Problemen, die dabei entstehen können. Es kann jedoch auch eine **interne** Fassung der Testbeschreibung geben**,** die im Allgemeinen streng vertraulich ist und die zusätzliche Hinweise für die Testautoren enthält.

Die interne Testbeschreibung sollte Hinweise zur Auswahl und Präsentation der Materialien enthalten, um die Testautoren davor zu bewahren, ihre eigenen, möglicherweise falschen Vorstellungen von Angemessenheit umzusetzen und dadurch Zeit zu verlieren. Die Hinweise könnten folgendermaßen aussehen:

i) Hinweise zur Auswahl der Texte

Diese könnten folgende Punkte beinhalten:

- die besten Textquellen (z. B. Artikel aus seriösen Zeitungen, Broschüren)
- Textquellen, die vermutlich weniger geeignet sind (z. B. Stadtteilzeitungen)
- eine allgemeine Warnung, kulturelle Voreingenommenheit zu vermeiden
- eine Liste von Gründen, warum Texte in der Vergangenheit abgelehnt wurden

Gründe für eine Ablehnung der Texte sind:

- es wird zu viel landeskundliches Hintergrundwissen vorausgesetzt
- unpassende Themen, wie z. B. Krieg, Tod, Politik und Glaubensfragen, die einige Kandidaten beleidigen oder sie emotional belasten könnten
- Themen, die außerhalb des Erfahrungsbereichs der Kandidaten liegen (z. B. aufgrund ihres Alters)
- zu hohe Anforderungen an den Wortschatz oder die Vorstellungskraft
- technische oder stilistische Fehler oder Eigentümlichkeiten
- schlecht redigierter Originaltext

Es ist auch denkbar, eine Liste der Themen zur Verfügung zu stellen, die so umfassend durch frühere Texte abgedeckt sind, dass keine weiteren Texte benötigt werden.

ii) Hinweise zur äußeren Form

Diese können z. B. die folgenden Punkte beinhalten:

- ob Texte mit zweifachem Zeilenabstand getippt werden sollen
- was in der Kopfzeile einer jeden Seite stehen soll
- ob Fotokopien der Originaltexte beigefügt werden sollen
- welche Angaben zur Textquelle gemacht werden sollen (z. B. Erscheinungsdatum)

iii) Detaillierte Hinweise zu jeder Aufgabe

Dies kann durch ein Beispiel verdeutlicht werden. Aufgabe ist ein modifizierter Cloze-Test, eher fokussiert auf Grammatik als auf Wortschatz:

Folgende Hinweise werden gegeben:

- Es wird ein authentischer Text von ca. 200 Wörtern benötigt. Er sollte eine kurze Überschrift haben. Die Betonung liegt auf einzelnen Strukturwörtern. Es sollte wenig unbekannter Wortschatz vorkommen.

- Es sollten mindestens 16 Items geschrieben werden; möglichst mehr, um nach der Erprobung eine Auswahl zu haben. Das erste Item wird das Beispiel und erhält die Nummer 0. Die Items sollen Präpositionen, Pronomen, Abtönungspartikel, Hilfsverben, Konjunktionen usw. testen und sich gleichmäßig über den Text verteilen. Es sollte darauf geachtet werden, dass die falsche Beantwortung eines Items nicht automatisch dazu führt, dass ein anderes falsch gelöst wird (Abhängigkeit der Items).

- Es empfiehlt sich im Allgemeinen nicht, die Lücke an den Anfang des Satzes zu setzen oder ein zusammengesetztes Wort als Lücke zu wählen, da dies die Kandidaten verwirren kann, weil sie nicht wissen, ob dies als ein Wort oder zwei Wörter gilt. Eine Lücke, ohne die der Satz auch vollständig wäre (z. B. „sehr“), sollte vermieden werden. Außerdem sollten Items (Lücken), die sehr ungewöhnliche oder eigentümliche Strukturen betreffen, vermieden werden.

- Die standardisierten Anweisungen im Test werden mitgeliefert. Ihnen folgt eine Checkliste, mit der die Materialien überprüft werden können, bevor man sie einreicht:

  Text:
  Ist das Thema leicht zugänglich/kulturell akzeptabel usw.?
  Hat der Text einen angemessenen Schwierigkeitsgrad?
  Ist der Text für eine Aufgabe mit Fokus auf den Strukturen angemessen?
  Ist der Text lang genug, um 16 Items zu ermöglichen?
  Hat der Text eine angemessene Überschrift?

  Items:
  Wurde die geforderte Anzahl an Items geschrieben?
  Verteilen sich die Items gleichmäßig über den Text?
  Gibt es eine gute Verteilung von Pronomen, Konjunktionen usw.?
  Wurde geprüft, ob sich alle Items auf die Strukturen beziehen?
  Wurde sichergestellt, dass es keine Itemabhängigkeit gibt?
  Wurden ein oder zwei zusätzliche Items geschrieben?
  Wurden Items mit ungewöhnlichen Strukturen vermieden?

  Anweisungen und Lösungsschlüssel:
  Wurde die Anweisung geprüft?
  Wurde ein Beispiel (0) gegeben?
  Wurde ein umfassender Lösungsschlüssel geschrieben?

Der Checkliste folgen einige Beispiele von abgelehnten Texten mit Hinweisen, warum diese Texte und Items ungeeignet sind.

Nachdem der Testautor alle zur Verfügung gestellten Informationen und Hinweise verarbeitet hat, muss er nun die Items erstellen und die Abgabefrist einhalten. Testautoren, die damit rechnen, Arbeitsaufträge über textbasierte Items zu bekommen, legen sich oft eine Textsammlung aus geeigneten Textquellen an und brauchen daraus nur auszuwählen, wenn der Arbeitsauftrag kommt. Um bestimmte Items schreiben zu können, ist es für den Testautor eine Hilfe, ein Wörterbuch oder einen Thesaurus zurate ziehen zu können.

Viele Testautoren halten es für sinnvoll, ihre Materialien mit einem Muttersprachler zu erproben, der nicht an der Testerstellung beteiligt ist. Dies hilft, Fehler zu entdecken, wie Tippfehler, unklare Anweisungen, falsche Lösungsschlüssel, Items, die sehr schwer zu lösen sind oder bei denen es mehrere richtige Lösungen gibt.

Bevor die Testautoren ihre Materialien an den Testentwickler schicken, sollten sie sicherstellen, dass sie sich Kopien gemacht haben. Wenn Artikel aus Zeitschriften oder Zeitungen verwendet werden und die Originale mit eingereicht werden müssen, sollten auch davon Kopien angefertigt und einbehalten werden.

b) Begutachtung und Redaktion

Wenn alle beauftragten Testautoren ihre Materialien eingesendet haben, kann als Erstes entschieden werden, welches Material in die Bearbeitung gehen kann und welches sofort abgelehnt werden sollte. Dies kann entweder durch Kollegen der Prüfungsinstitution oder durch externe Berater geschehen.

In dieser Phase können Texte, die aus den oben genannten Gründen eindeutig ungeeignet sind, abgelehnt werden. Wenn der Auftrag lautete, Texte ohne Items einzureichen, können die Testautoren nun gebeten werden, Items zu den Texten zu schreiben, die bei der Begutachtung angenommen wurden. Für Testautoren, die ihre Texte ohne Items einreichen, ist es immer ratsam, mindestens in einer vorläufigen Fassung Items dazu zu schreiben, um sie zügig nachliefern zu können, sobald die Texte angenommen wurden.

Externe Testautoren sind normalerweise nicht an der Begutachtung beteiligt, häufig werden sie aber gebeten, an Testredaktionssitzungen teilzunehmen, in denen das in Auftrag gegebene Material verabschiedet wird. Sie bekommen dafür ein Honorar, das die Dauer der Sitzung und die Zeit abdeckt, die sie für die sorgfältige Lektüre der Testmaterialien aufgewendet haben. Der Testentwickler muss entscheiden:

- welche Gruppen er während der Redaktionssitzung bildet
- welche Materialien jede Gruppe bearbeiten soll

Die Materialien für die Redaktionssitzung sollten den Teilnehmern im Voraus geschickt werden, damit jeder Zeit hat, sich vor der Sitzung die Unterlagen sorgfältig anzusehen. Die Texte sollten nicht nur gelesen werden, um ihre Länge, die Angemessenheit des Themas und des Stils sowie die sprachliche Schwierigkeit zu überprüfen. Es empfiehlt sich auch, den Lösungsschlüssel zu überprüfen (falls er mitgeliefert wird) und die Items so zu lösen, als würde man die Prüfung ablegen. Dadurch werden verschiedene Probleme ersichtlich:

- ob es z. B. Multiple-Choice-Items gibt, bei denen mehr als eine richtige Antwort möglich ist
- ob es z. B. Multiple-Choice-Items gibt, bei denen die Lösung unklar oder schlecht formuliert wurde
- ob es Distraktoren gibt, die so unplausibel sind, dass sie vermutlich von niemandem gewählt werden und
- ob es Items gibt, die selbst für einen Muttersprachler schwer oder unklar wären

Diese Materialien sollten stets vertraulich behandelt werden.

Neue Testautoren, die an Redaktionssitzungen teilnehmen, können dadurch geschult werden, dass sie in Gruppen mit erfahrenen Teilnehmern arbeiten. Bei Redaktionssitzungen erschwert eine Gruppe mit mehr als vier oder fünf Teilnehmern die Arbeit, während eine Gruppe unter drei Teilnehmern vermutlich nicht genug verschiedene Gesichtspunkte einbringt.

In jeder Gruppe muss ein Teilnehmer Protokoll über alle Entscheidungen hinsichtlich der Materialien führen, damit jede Veränderung, die während der Sitzung vorgenommen wurde, genau festgehalten wird. Es ist ungewöhnlich, dass die Materialien so akzeptiert werden, wie sie eingereicht wurden. Im Gegenteil, es ist üblich, dass die angenommenen Materialien während der Redaktionssitzung erheblich überarbeitet werden. Es entstehen oft heftige Diskussionen über die Materialien und besonders bei Texten auf einem hohen Niveau gibt es unterschiedliche Meinungen über deren Angemessenheit, die alle begründet werden können. Aus diesem Grund ist es hilfreich, wenn der Testentwickler oder eine andere Person mit einer gewissen Autorität endgültige Entscheidungen fällen und beschließen kann, wann die Diskussion abgeschlossen werden soll.

Es gibt unterschiedliche Ansichten darüber, ob Testautoren in Redaktionssitzungen ihre eigenen Materialien überarbeiten sollten. Manche Testentwickler versuchen, dies zu vermeiden, aber wenn nur wenige Personen an der Überarbeitung beteiligt sind, muss vermutlich jeder seine eigenen Materialien überarbeiten. Wegen der unmittelbaren Rückmeldung und der Möglichkeit, in Zukunft die gleichen Fehler zu vermeiden, kann es für den Testautor besser sein, die Meinungen der anderen direkt zu hören, obgleich es schwierig ist, dabei zu sein, wenn Materialien, an denen man lange gearbeitet hat, kritisiert oder abgelehnt werden.

Am Ende der Redaktionssitzung sollte kein Zweifel (auch kein zukünftiger) darüber bestehen, welche Veränderungen verabschiedet wurden. Eine klare Niederschrift der Veränderungen, die an den angenommenen Materialien vorgenommen wurden, muss aufbewahrt werden. Manche eingereichten Materialien zeigen Potenzial, setzen aber eine umfangreiche Überarbeitung voraus, die in einer Redaktionssitzung nicht zu leisten wäre. Sie könnten dazu an den Autor zurückgegeben werden (normalerweise ohne zusätzliches Honorar) oder an einen anderen Testautor mit mehr Erfahrung. Es sollte deutlich gesagt werden, ob Materialien, die aus bestimmten Gründen abgelehnt wurden, dem Testautor zurückgeschickt werden, der sie dann anderweitig verwenden kann, z. B. im Unterricht, oder ob die Materialien Eigentum der Prüfungsinstitution bleiben. Es könnte sein, dass der Testentwickler Beispiele für abgelehnte Materialien sammeln möchte, um sie in Trainingsseminaren zu verwenden.

Nach der Redaktionssitzung werden die gebrauchten und nicht gebrauchten Kopien der Materialien vernichtet. Die überarbeiteten Fassungen des angenommenen Materials werden vom Testentwickler aufbewahrt und in Vorbereitung der Erprobung neu abgetippt. Dabei werden die Namen der Autoren und Hinweise auf Textquellen weggelassen.

Testautoren haben ein Anrecht, vom Testentwickler eine Rückmeldung über die abgelehnten Texte oder Items zu bekommen, besonders wenn sie nicht zur Redaktionssitzung eingeladen wurden oder ihr eigenes Material nicht überarbeitet haben. Dies hilft ihnen, nicht diesselben Fehler zu wiederholen, wenn sie zukünftig neues Material einreichen.

Zu diesem Zeitpunkt endet für gewöhnlich die Mitarbeit der Testautoren im Produktionsprozess der Prüfung.

c) Erprobung

Nach der Phase der Itemerstellung, der Begutachtung und der Überarbeitung ist die Erprobung eine weitere, objektive Überprüfung, ob ein Item gut genug ist, um in die endgültige Prüfung aufgenommen zu werden.

Normalerweise werden die objektiv zu bewertenden Itemtypen wie Multiple-Choice oder Lückentexte mit einer größeren Probandengruppe erprobt (engl. *pretested*). Dabei werden die einzelnen Items überprüft, nicht der Test als Ganzes. Deshalb muss die Fassung, die in die Erprobung geht, nicht notwendigerweise wie die endgültige Prüfung aussehen, für die die Items geschrieben wurden, weder was die Länge noch was die Zusammenstellung betrifft. So kann man z. B. für eine Prüfung, die aus 20 einzelnen Multiple-Choice-Items und drei Texten mit jeweils

fünf Multiple-Choice-Items besteht, eine Erprobungsfassung mit 40 einzelnen Multiple-Choice-Items, gefolgt von einem Text mit fünf Items, zusammenstellen.

Erprobungen werden unter Prüfungsbedingungen mit Lernern der Fremdsprache durchgeführt, deren Sprachniveau die Lehrer als angemessen einschätzen, um die Prüfung ablegen zu können. Die Lerner bekommen durch die Erprobung Übung für die Prüfung und eine Rückmeldung über ihre Leistungen; die Prüfungsorganisation erhält die Daten aus den statistischen Analysen, nachdem die Antwortbögen in der üblichen Weise erfasst und ausgewertet wurden. Damit diese Daten aussagekräftig sind, geht man normalerweise von einer Erprobungspopulation von mindestens 100 Lernern aus.

Die subjektiv bewerteten Teile des Tests, wie z. B. Aufsätze, können nicht auf die gleiche Art und Weise erprobt werden wie Items, für die es nur eine richtige Antwort gibt. Jedoch können auch diese Aufgabentypen überprüft werden, bevor sie in der endgültigen Prüfung eingesetzt werden. Sie können mit **kleineren Probandengruppen erprobt** werden (engl. *trialling*), wiederum mit freiwilligen Teilnehmern, die das richtige sprachliche Niveau haben. Die Arbeiten der Lerner werden von Beurteilern, die mit der Prüfung vertraut sind, nach den Beurteilungskriterien der Prüfung bewertet. Dadurch kann man sehen, ob die Aufgabe von den Lernern verstanden wurde, ob sie ihrem Alter und ihrem Erfahrungshintergrund angemessen war, ob den Lernern genügend Informationen gegeben wurden, um die Aufgabe angemessen auszuführen und ob die Aufgabe ihnen erlaubte, das Spektrum an Wortschatz und Strukturen zu zeigen, das von einem Kandidaten in einer Prüfung auf diesem Niveau erwartet wird.

d) Statistische Analysen der Items

Statistische Analysen der Lösungen der Items geben dem Testentwickler sehr nützliche Informationen über die Qualität der Items. Sie können dazu beitragen, den endgültigen Test von ungeeigneten oder fehlerhaften Items freizuhalten. Man muss jedoch bedenken, dass auch ein ungeeignetes Item akzeptable statistische Werte aufweisen kann. Deshalb sollte man die Ergebnisse der statistischen Analysen nur als einen Faktor unter anderen sehen, die bestimmen, welche Materialien in die endgültige Prüfung aufgenommen werden.

Die Daten, die während der Erprobung erhoben wurden, können entweder mit der klassischen Itemanalyse oder der Rasch-Analyse ausgewertet werden. Für die klassische statistische Analyse wird das Computerprogramm *Iteman* aus *MicroCAT* verwendet. Diese Art von Analyse gibt Auskunft über die Qualität eines jeden Items. Man erfährt Folgendes:

i) Schwierigkeit der Items

Man muss wissen, wie leicht oder schwierig ein Item ist, um sicherzustellen, dass die Testmaterialien für die Kandidaten den richtigen Schwierigkeitsgrad haben. Die Schwierigkeit wird ausgedrückt als **Anteil der korrekten Antworten an der Gesamtzahl der Kandidatenantworten** bei einem Item (vgl. die Spalte ‚**Prop. Correct**', *proportion correct* in Abbildung 4). Die angemessene Schwierigkeit des Tests liegt in der Mitte der Schwierigkeitsskala, kann aber auch zwischen Werten von .33 und .67 liegen oder zwischen .20 und .80. Dies richtet sich nach dem Zweck, zu dem der Test eingesetzt wird, und kann von Test zu Test variieren. Bei einem Sprachtest am Ende eines Kurses können andere Schwierigkeitswerte als bei einem Eignungstest erwartet werden.

Ein Test sollte einige Items enthalten, die auf den beiden Extremen der Skala angesiedelt sind. Häufig werden besonders am Anfang des Tests einige leichte Items aufgenommen, damit die Kandidaten sich auf den Test einstellen können. Manchmal werden diese leichten Items bei der Ermittlung des Ergebnisses nicht mitgezählt.

Items, deren Werte außerhalb der für den Test festgesetzten Streuung liegen, werden nach der Erprobung verworfen, aber wenn es eine Itembank gibt, können sie dort aufgenommen werden und in einem Test auf einem anderen Niveau eingesetzt werden.

ii) Trennschärfe der Items

Dieser statistische Wert zeigt an, wie gut das Item die schwachen von den starken Kandidaten trennt. Mehr Lerner mit einem hohen Gesamtergebnis (engl. *score*) im Test sollten ein Item richtig lösen als Lerner mit einem niedrigen Gesamtergebnis. Zwei Methoden werden verwendet, um die Trennschärfe zu errechnen: der Trennschärfeindex und die punkt-biseriale Korrelation (vgl. die Spalten ‚**Disc. Index'**, *discrimination index* und ‚**Point Biser'**, *point biserial correlation* in Abbildung 4).

Trennschärfeindex

Nachdem ein Test mit einer Anzahl von Kandidaten erprobt wurde, werden die Kandidaten nach ihren Gesamtergebnissen sortiert und in eine Rangfolge gebracht. Zwei Gruppen werden dann aus der Analyse herausgenommen: die oberen 27% der Kandidaten, die sehr leistungsstarke Gruppe, und die unteren 27% der Kandidaten, die leistungsschwache Gruppe. Auf der folgenden Seite finden Sie einen Auszug aus einer statistischen Analyse mit dem Programm Iteman aus MicroCAT (Abbildung 4).

MicroCAT (™) Testing System
Copyright © 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation

Item and Test Analysis Program – ITEMAN (tm) Version 3.50

Item analysis for data from file C:\ITEMAN\13002D93.D93 Time: 15.59

Item Statistics					Alternative Statistics					
Seq No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
8	2-1	.38	.52	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					C	.12	.11	.12	-.01	
					D	.49	.74	.23	-.44	
					Other	.01	.00	.00	-.11	
9	2-2	.71	.42	.42	A	.07	.11	.01	-.16	
					B	.11	.18	.04	-.22	
					C	.10	.16	.00	-.22	
					D	.71	.53	.95	.42	*
					Other	.01	.00	.00	-.13	
10	2-3	.68	.56	.56	A	.68	.39	.96	.56	*
					B	.21	.36	.04	-.37	
					C	.03	.08	.00	-.24	
					D	.07	.14	.00	-.22	
					Other	.01	.00	.00	-.13	
11	2-4	.57	.49	.49	A	.18	.28	.08	-.27	
					B	.15	.19	.09	-.12	
					C	.08	.16	.01	-.31	
					D	.57	.33	.83	.49	*
					Other	.01	.00	.00	-.13	
12	2-5	.61	.63	.54	A	.09	.18	.00	-.22	
					B	.20	.28	.03	-.27	
					C	.61	.32	.96	.54	*
					D	.09	.18	.01	-.28	
					Other	.02	.00	.00	-.09	
13	2-6	.81	.35	.48	A	.11	.20	.04	-.29	
					B	.01	.03	.00	-.11	
					C	.81	.61	.96	.48	*
					D	.07	.17	.00	-.34	
					Other	.00	.00	.00		
14	3-1	.93	.19	.39	A	.93	.81	1.00	.39	*
					B	.07	.18	.00	-.39	
					Other	.01	.00	.00	-.03	

Abbildung 4: Itemanalyse mit MicroCAT (Iteman)

Die Anzahl der Kandidaten in den beiden Gruppen (der leistungsstarken und der leistungsschwachen Gruppe) werden als N ausgewiesen. Man zählt die Kandidaten in jeder Gruppe, die das Item richtig gelöst haben. Dies ergibt:

n_H = die Anzahl der Kandidaten in der leistungsstarken Gruppe, die das Item richtig gelöst haben

n_L = die Anzahl der Kandidaten in der leistungsschwachen Gruppe, die das Item richtig gelöst haben

Der Trennschärfeindex d_i kann dann folgendermaßen berechnet werden:

$$d_i = \frac{n_H - n_L}{N}$$

Der Trennschärfeindex d_i kann jeden Wert zwischen -1 und +1 annehmen.

Ein Trennschärfeindex von +1 bedeutet, dass alle „guten" Kandidaten das Item richtig gelöst und alle „schlechten" Kandidaten das Item falsch gelöst haben. Ein Trennschärfeindex von -1 bedeutet, dass alle „guten" Kandidaten das Item falsch gelöst und alle „schlechten" Kandidaten das Item richtig gelöst haben.

Items mit einem d_i von 0.30 oder höher werden im Allgemeinen als angemessen für die betreffende Gruppe angesehen. Es ist zu beachten, dass der Trennschärfeindex abhängig ist von der Fähigkeit der jeweiligen Probandengruppe.

Punkt-biseriale Korrelation

Die punkt-biseriale Korrelation r_{pb} wird durch folgende Formel berechnet:

$$r_{pb} = \frac{\overline{x_p} - \overline{x_q}}{s_x}\sqrt{pq}$$

wobei $\overline{x_p}$ der Mittelwert des Gesamtpunktwerts (Gesamtscore) derjenigen Kandidaten ist, die dieses Item richtig gelöst haben, und

$\overline{x_q}$ der Mittelwert des Gesamtscores derjenigen Kandidaten ist, die dieses Item falsch gelöst haben.

p ist der Anteil aus allen Kandidaten, die das Item richtig gelöst haben, und

q der Anteil aus allen Kandidaten, die das Item falsch gelöst haben.

s_x ist die Standardabweichung bei den Gesamtscores aller Kandidaten.

Im Allgemeinen werden Items mit einer punkt-biserialen Korrelation von 0.30 und größer als angemessen angesehen.

Erhält man negative punkt-biseriale Korrelationen, bedeutet dies, dass leistungsstarke Kandidaten das Item falsch gelöst haben. Dies kann bedeuten, dass neben der Lösung auch eine der anderen Optionen eine richtige Antwort sein kann. Man nennt diese Option dann einen positiven Distraktor. Ein solches Item sollte nicht in einem Test eingesetzt werden, man kann es aber eventuell überarbeiten und erneut erproben.

iii) Angaben zu den Distraktoren

Die statistische Analyse von Multiple-Choice-Items zeigt, ob die Distraktoren angemessen funktionieren, mit anderen Worten, ob jeder Distraktor plausibel genug ist, um von einigen der Kandidaten als richtige Antwort angesehen zu werden. Er sollte aber nicht so nah an der tatsächlichen Lösung liegen, dass mehr Kandidaten den Distraktor anstelle der richtigen Lösung wählen.

Auf dem Computerausdruck wird bei jedem Distraktor der Anteil der Kandidaten ausgewiesen, die diesen Distraktor gewählt haben, und zwar in der Spalte Prop. Total. Um einige Beispiele zu geben:

Eine „Prop. Total"-Spalte zeigt Folgendes:	A	.15
	B	.10
	C	.63 *
	D	.12

Die richtige Lösung ist C (im MicroCAT-Ausdruck mit * gekennzeichnet). Es handelt sich um ein Item, bei dem die richtige Lösung und alle Distraktoren angemessen funktionieren.

Eine andere „Prop. Total"-Spalte zeigt Folgendes:	A	.95 *
	B	.04
	C	.01
	D	.00

Die richtige Lösung ist A. Es handelt sich um ein Item, das fast alle Kandidaten richtig beantwortet haben und bei dem ein Distraktor (D) so schwach war, dass niemand ihn gewählt hat.

Im Ausdruck kann man außerdem die Spalten **„Seq. No."** und **„Scale-item"** finden. Sie beziehen sich auf die **durchlaufende Nummerierung aller erprobten Items** (Seq. No.) und die **Nummer des Items innerhalb eines Subtests** (Scale-item, wobei die erste Nummer der Subtest ist, die zweite das Item in diesem Subtest).

Man kann auch Informationen darüber erhalten, wie der Test (d. h. die Erprobungsfassung) als Ganzes mit dieser bestimmten Kandidatengruppe funktioniert. Ein Beispiel für einen solchen Ausdruck findet sich in Abbildung 5. Die statistischen Begriffe, die in den Spalten „scale statistics" aufgeführt sind, bedeuten Folgendes:

N of Items:	Anzahl der Items in der Analyse.
N of Examinees:	Anzahl der Kandidaten in der Analyse.
Mean:	Mittelwert. Bei dichotomen Items: die durchschnittliche Anzahl richtig gelöster Items. Bei trichotomen Items: der mittlere Wert des Scores der Kandidaten im jeweiligen Subtest.
Variance:	Varianz. Maß für die Streuung der Gesamtpunktwerte (*engl. scores*) um den Mittelwert.

Std. Dev.: Die Standardabweichung (engl. *standard deviation)* ist die Quadratwurzel der Varianz.

Skew: Schiefe. Kennwert für das Ausmaß, in dem eine Verteilung von der Symmetrie abweicht.

Kurtosis: Exzess. Kennwert für das Ausmaß, in dem eine Verteilung flacher oder spitzer als eine Normalverteilung ist.

Minimum: Minimum. Niedrigster Gesamtpunktwert der Kandidaten.

Maximum: Maximum. Höchster Gesamtpunktwert der Kandidaten.

Median: Median. Mittlerer Gesamtpunktwert der Kandidaten.

Alpha: Alpha Koeffizient. Eine Reliabilitätsschätzung auf der Basis der Messung der inneren Konsistenz eines Tests. Der Wert von Alpha reicht von 0 bis 1. Idealerweise sollte er so nahe an 1 sein wie möglich.

SEM: Standardmessfehler (engl. *standard error of measurement).* Zeigt den vermutlichen „Fehler", d. h., er ist ein Indikator für die Ungenauigkeit der Messung der Items.

$$SEM = SD_x \sqrt{(1 - R_x)}$$

SEM = Standardmessfehler
SD = Standardabweichung
R (x) = Reliabilität des Tests[1]

Bei Vorliegen einer Normalverteilung liegt mit einer Wahrscheinlichkeit von 68% der beobachtete Score innerhalb einer Standardabweichung ($\pm$1 SEM) und mit einer Wahrscheinlichkeit von 95% innerhalb von zwei Standardabweichungen ($\pm$2 SEM) links und rechts vom Mittelwert.

Beispiel: Ein Kandidat erreicht 67 Punkte in einem Test, der eine Standardabweichung von 9 und einen Reliabilitätskoeffizienten von 0.9 hat:

$$SEM = 9 \sqrt{(1 - 0.9)} = 2.8$$

Mit einer Wahrscheinlichkeit von 68% liegt der wahre Score des Kandidaten zwischen 64.2 und 69.8 Punkten. Und mit einer Wahrscheinlichkeit von 95% zwischen 61.4 und 72.6 Punkten.

Mean P: Mittlere Schwierigkeit. Durchschnittliche Lösungsrate (gilt nur für dichotome Items).

[1] Anmerkung (TestDaF-Institut): „x" steht hier für einen beliebigen Test.

Mean Item-Tot.: Mittlerer punkt-biserialer Wert für den jeweiligen Subtest (gilt nur für dichotome Items).

Mean Biserial: Mittlere biseriale Korrelation für den jeweiligen Subtest.

Max Score (Low): Der höchste Score, den ein Kandidat in der leistungsschwachen Gruppe (die unteren 27%) erreichen konnte.

N (Low Group): Die Anzahl der Kandidaten in der leistungsschwachen Gruppe. Dies sind ca. 27% aller Kandidaten.

Min Score (High): Der niedrigste Score, den ein Kandidat in der leistungsstarken Gruppe erreicht hat.

N (High Group) Die Anzahl der Kandidaten in der leistungsstarken Gruppe. Dies sind ca. 27% aller Kandidaten.

Die Zusammenfassung einer Analyse mit MicroCAT (Iteman) finden Sie auf der nächsten Seite (Abbildung 5).

MicroCAT (™) Testing System
Copyright © 1982, 1984, 1986,1988, 1993 by Assessment Systems Corporation

Item and Test Analysis Program – ITEMAN (™) Version 3.50

Item analysis for data file C:\ITEMAN\13001D93.DAT

Time:15.59

Missing-data option: Compute statistics on all available item responses

There were 270 examinees in the data file.

Scale Statistics

Scale:	1	2	3	4	5	6	7
……	……	……	……	……	……	……..	……
…….	…	….	…..	…..	….	.	…
N of items	5	10	10	10	6	5	15
N of Examinees	270	270	270	270	270	270	270
Mean	3.230	6.633	8.422	8.163	3.778	1.959	12.807
Variance	0.725	3.321	1.755	2.588	1.751	1.321	3.259
Std. Dev.	0.851	1.822	1.325	1.609	1.323	1.149	1.805
Skew	0.047	-0.348	-0.361	-0.709	-0.288	0.402	-0.622
Kurtosis	-0.491	-0.202	3.043	-0.148	-0.325	-0.153	-0.292
Minimum	1.000	1.000	2.000	3.000	0.000	0.000	7.000
Maximum	5.000	10.000	10.000	10.000	6.000	5.000	15.000
Median	3.000	7.000	9.000	8.000	4.000	2.000	13.000
Alpha	0.091	0.431	0.318	0.499	0.371	0.407	0.570
SEM	0.812	1.375	1.094	1.138	1.050	0.885	1.183
Mean P	0.646	0.663	0.842	0.816	0.630	0.392	0.854
Mean Item – Tot.	0.428	0.406	0.378	0.415	0.493	0.541	0.342
Mean Biserial	0.676	0.547	0.602	0.621	0.662	0.753	0.590
Max Score (Low)	3	6	8	7	3	1	12
N(Low Group	168	116	115	89	115	103	105
Min Score (High)	4	8	9	9	5	3	14
N (High Group)	102	85	155	132	89	84	109

Abbildung 5: MicroCAT-Analyse (Iteman)

e) Kalibrierung der Items

Die oben beschriebenen klassischen Itemanalysen sind zwar nützlich, ihre Informationen über das Funktionieren der Items sind jedoch abhängig von der Probandengruppe, die an der Erprobung des Tests teilgenommen hat. Es können allgemeingültige Aussagen über die Schwierigkeit des Tests abgeleitet werden, besonders wenn viele Informationen über die Erprobungspopulation vorliegen. Es ist aber schwierig, genaue Vergleiche zwischen Items zu ziehen, die unter verschiedenen Umständen erprobt wurden. Deshalb wurde die Rasch-Analyse als ein zusätzliches Analyseverfahren entwickelt, das die klassische Analyse ergänzt. Man kann eine Schwierigkeitsskala erstellen, indem man bei jeder Erprobung Items mit schon bekannten statistischen Werten in den Test aufnimmt und dann die Schwierigkeit der neuen Items mit den bekannten Schwierigkeitswerten vergleicht. Diese Items mit bekannten Werten werden in jede Erprobungsfassung eines Tests aufgenommen. Im Grunde besteht die Itemkalibrierung darin, Items und Kandidaten aus verschiedenen Erprobungen auf einer gemeinsamen Skala abzubilden.

Um eine Kalibrierung zu ermöglichen, muss erst eine Schwierigkeitsskala für einen Echteinsatz der jeweiligen Prüfung entwickelt werden. Danach werden **Ankeritems** aus dieser Prüfung ausgewählt und als gemeinsame Items in jede Erprobungsfassung aufgenommen. Die Schwierigkeit der neuen Items wird berechnet, indem man sie mit den Ankeritems vergleicht und auf der gemeinsamen Schwierigkeitsskala abbildet. Wenn eine Prüfung ausschließlich aus kalibrierten Items zusammengestellt wird, kann schon vor dem Echteinsatz der Prüfung ermittelt werden, welche Scores (Gesamtpunktwert im Test) welcher Note oder Einstufung entsprechen (siehe Abbildung 6).

Die Rasch-Analyse der Daten ermöglicht den Aufbau einer **computerbasierten Itembank** und die systematische Verwaltung der Items. Dies wäre die nächste Phase im Testerstellungprozess. Wenn diese Möglichkeit nicht besteht, bewahrt der Testentwickler die erprobten Items auf (die nicht verworfen wurden oder revisionsbedürftig sind), um sie bei der Zusammenstellung des endgültigen Tests zu verwenden.

i) Erstellen einer Schwierigkeitsskala

Institutionen, die in regelmäßigen Abständen Tests anbieten, müssen in jedem Fall sicherstellen, dass die Bestehensgrenzen, die sie auf der Bewertungsskala setzen (z. B. zwischen „bestanden“ und „nicht bestanden“ oder zwischen den Noten „sehr gut“ und „gut“), von Testdurchführung zu Testdurchführung gleich bleiben. Traditionellerweise wird dies mittels eines Notengebungsverfahrens geleistet, das zunächst ermittelt, ob sich die Kandidatenpopulation in irgendeiner Weise seit der letzten Prüfung verändert hat, und das dann angemessene Bestehensgrenzen festsetzt, wobei die Schwierigkeit des Tests und die Fähigkeit der Kandidaten berücksichtigt werden.

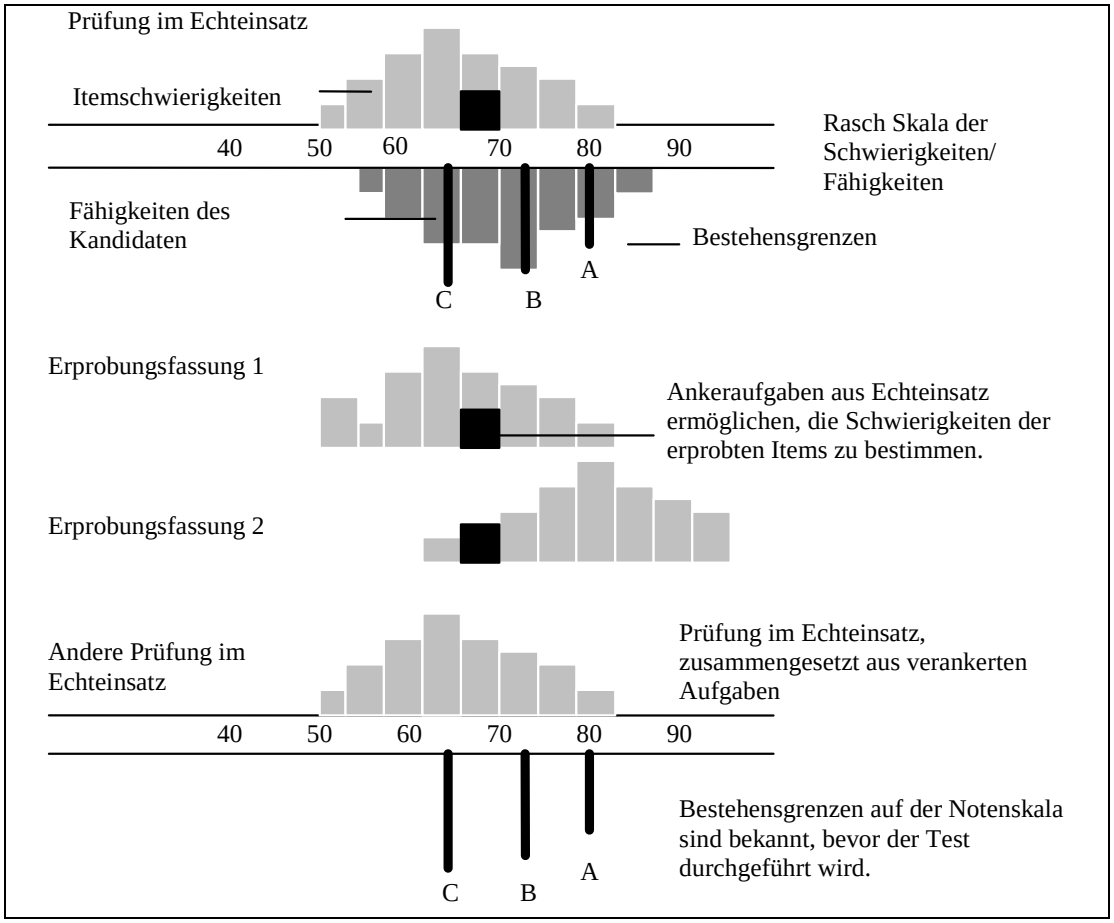

Abbildung 6: Kalibrierung der Items

Normalerweise müssen die Schwierigkeitswerte der Tests im Voraus bekannt sein, um sicherzustellen, dass alle folgenden späteren Versionen des Tests äquivalent und deshalb fair gegenüber den Kandidaten sind, die den Test zu einem späteren Zeitpunkt ablegen. Selbst wenn das oben erwähnte Notengebungsverfahren angewendet wird, ist es trotzdem noch wünschenswert, dass alle Tests ein ähnliches statistisches Profil aufweisen, denn Institutionen möchten sicher sein, dass Kandidaten, die eine Prüfung wiederholen, damit vermutlich ihre Endnoten verbessern. Dies setzt natürlich voraus, dass sie in der Zwischenzeit einen Kurs besucht haben und nicht durch ungünstige Umstände wie Krankheit oder Nervosität beeinträchtigt werden. Aus diesem Grund ist es wichtig, die Fähigkeiten einer Testpopulation im Vergleich zu einer repräsentativen Auswahl von Items zu kennen. Dazu werden Daten von ca. 300 Kandidaten aus einer durchgeführten Prüfung mithilfe der Rasch-Analyse analysiert. Diese Analyse erlaubt die Einschätzung der Schwierigkeit aller Items durch einen Vergleich untereinander. Das schwierigste Item ist offensichtlich das, welches die wenigsten Kandidaten richtig lösen konnten usw. Der Unterschied zwischen diesem Verfahren und der klassischen Analyse besteht darin, dass die Rasch-Analyse die Items nach der Wahrscheinlichkeit anordnet, mit der die Kandidaten das Item richtig lösen werden. Sie tut dies, indem sie die Itemschwierigkeiten auf einer Wahrscheinlichkeitsskala ausweist, deren Messeinheiten *Logits* genannt werden. Das Wort *Logit* kommt von dem logarithmischen Verfahren, das angewendet wird, um die Itemschwierigkeit zu bestimmen. Die Skala hat einen Mittelwert von 0, und die Schwierigkeitswerte werden als positive Werte (schwerer) nach oben oder als negative Werte (leichter) nach unten weitergeführt. Wenn z. B. das Item 1 eine Schwierigkeit von 4.2 hat und das Item 2 eine Schwierigkeit von -0.2, kann man gut erkennen, dass das zweite Item leichter ist als das erste.

Die Schwierigkeit von Items kann nur im Zusammenhang mit der Fähigkeit von Kandidaten bestimmt werden. Deshalb weist die Rasch-Analyse die Schwierigkeiten der Items auf einer Skala so aus, dass sie sich untereinander vergleichen lassen und bildet die Fähigkeit der Kandidaten auf der gleichen Skala ab. Die fähigsten Kandidaten erhalten höhere Fähigkeitswerte; damit ist z. B. ein Kandidat mit einem Fähigkeitswert von 2.3 bei diesem Test fähiger als ein Kandidat mit einem Fähigkeitswert von -1.5 (vgl. Abbildung 7).

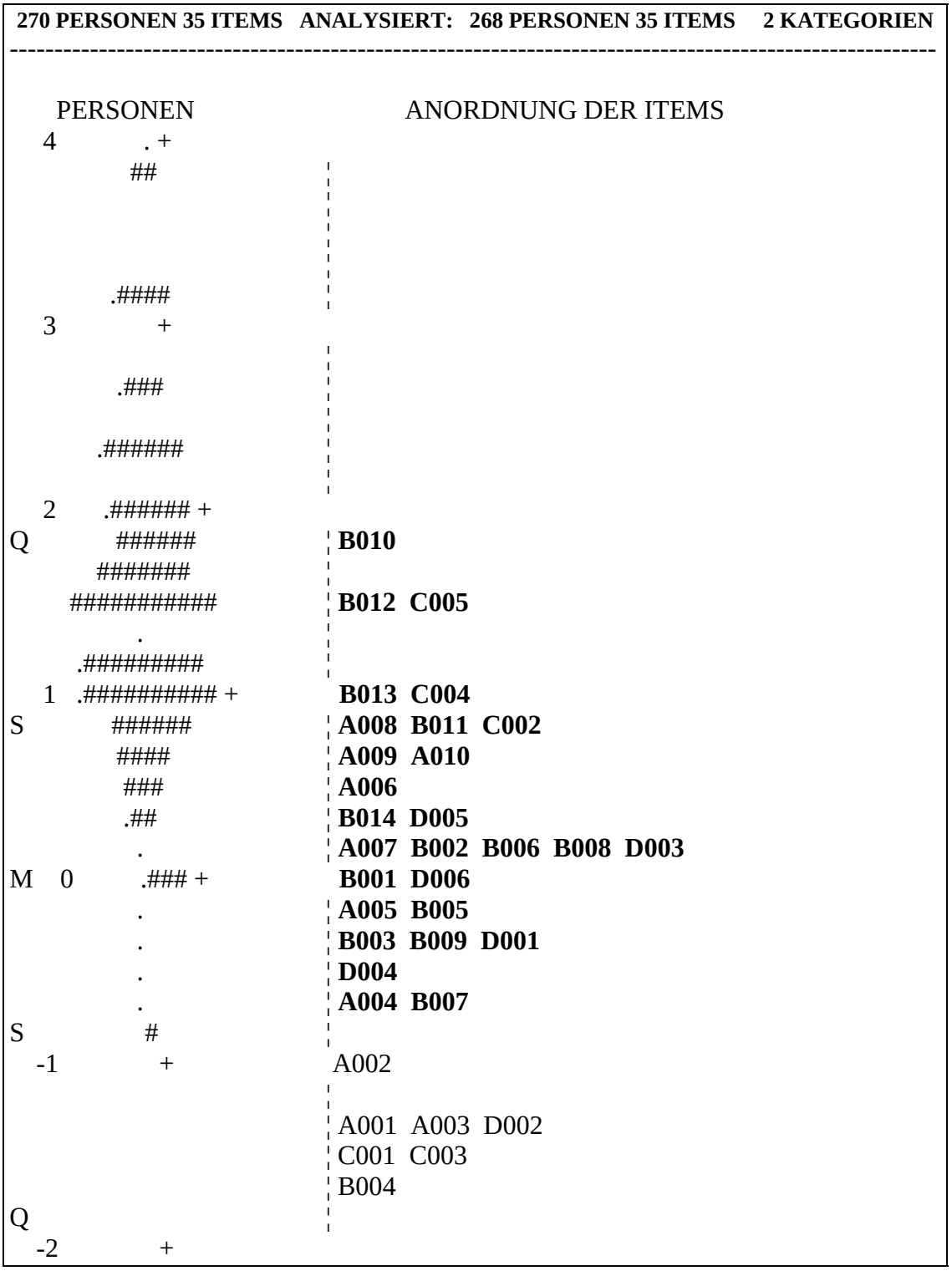

Abbildung 7: Eine Rasch-Skala

Die Abbildung zeigt rechts die Schwierigkeitswerte der Items im Vergleich zu den Fähigkeitswerten der Kandidaten links. Item B010 ist das schwierigste und Item B004 das leichteste Item. Die Buchstaben vor den Nummern beziehen sich auf die Itemgruppe in der Analyse, da man bei manchen Tests die Items vielleicht nach Gruppen oder Testteilen ausweisen möchte. In diesem Test, auf den sich die Analyse oben bezieht, waren einige Items für die Kandidaten, die den Test abgelegt haben, zu leicht. Die fettgedruckten Items haben angemessene Schwierigkeitswerte, hingegen sind alle Items mit einem *Logit* unter -1 zu leicht. Dies resultiert

daraus, dass der mittlere Fähigkeitswert der Kandidaten, die den Test abgelegt haben, bei 1.45 liegt; der mittlere Schwierigkeitswert der Items liegt stets bei 0. Dieser Unterschied von zwei *Logits* bedeutet, dass es für den durchschnittlichen Kandidaten eine Wahrscheinlichkeit von 80% gibt, diese Items richtig zu lösen, denn in dem Maße, wie der Fähigkeitswert höher liegt als die Itemschwierigkeit, steigt auch die Wahrscheinlichkeit, das Item richtig zu lösen. Abbildung 8 zeigt, wie man die Fähigkeitswerte interpretieren kann.

Logit-Differenz zwischen Fähigkeitswert und Itemkalibrierung	Lösungswahrscheinlichkeit bei einem dichotomen Item	*Logit*-Differenz zwischen Fähigkeitswert und Itemkalibrierung	Lösungswahrscheinlichkeit bei einem dichotomen Item
5.0	99%	-5.0	1%
4.6	99%	-4.6	1%
4.0	98%	-4.0	2%
3.0	95%	-3.0	5%
2.2	90%	-2.2	10%
2.0	88%	-2.0	12%
1.4	80%	-1.4	20%
1.1	75%	-1.1	25%
1.0	73%	-1.0	27%
0.8	70%	-0.8	30%
0.5	62%	-0.5	38%
0.4	60%	-0.4	40%
0.2	55%	-0.2	45%
0.1	52%	-0.1	48%
0	50%	0	50%

Abbildung 8: Umrechnungstabelle *Logit*-Wahrscheinlichkeit

Idealerweise sollten die Schwierigkeitswerte der Items den Fähigkeitswerten der Kandidaten entsprechen, damit der Test im Hinblick auf seine Schwierigkeit angemessen ist.

Wenn die Analyse beendet ist, die Schwierigkeitswerte zufriedenstellend und in Bezug auf die Teilnehmerpopulation bekannt sind, kann dieser Test als Fixpunkt verwendet werden, um weitere neue Tests mit ihm zu verankern. Dies geschieht dadurch, dass man Ankeraufgaben aus diesem Test auswählt.

ii) Verankerung

Im Allgemeinen werden den Erprobungsfassungen der Tests Ankeraufgaben beigefügt um sicherzustellen, dass alle neuen Testaufgaben auf einer gemeinsamen Schwierigkeitsskala abgebildet werden können. Dabei geht man davon aus, dass der ursprüngliche Test, d. h. der Fixpunkt, mit dem alle zukünftigen Tests verankert werden, geheim bleibt und nicht als Übungsmaterial verwendet wird oder Kandidaten davon nach der Prüfung Exemplare mitnehmen durften.

Die Verankerung geschieht folgendermaßen:

- Man wählt eine Gruppe von Items aus, die im Hinblick auf Schwierigkeit und Modellanpassung angemessen sind (im Deutschen wird dafür auch der englische Begriff *fit* gebraucht). Der *fit*-Index gibt die Anpassung der beobachteten Daten an das Modell wieder. In der Regel weicht die *fit*-Statistik nur unwesentlich von dem Mittelwert 1 ab.
- Man fügt diese Items allen Tests bei, die in die Erprobung gehen.
- Für die ausgewählten Items erstellt man eine Anker-Datei mit einer Liste der Schwierigkeitswerte ausgedrückt in *Logits*. Die Ankeraufgaben werden entsprechend ihrer Platzierung in der Erprobungsfassung nummeriert und als ASCII-Datei gespeichert.
- In der Rasch-Datei gibt man den Namen der Anker-Datei an (IAFILE = Dateiname, Laufwerk usw. a:/anchor.anc).[2]
- Man führt die Analyse durch. Die mittleren Schwierigkeitswerte liegen nicht mehr bei 0, da die Schwierigkeit der Items nun in Bezug auf die Fähigkeitswerte der ursprünglichen Prüfungspopulation geschätzt wird.

Einige Institutionen haben vielleicht nicht die Mittel, das oben genannte Verankerungsverfahren als regulären Bestandteil des Entwicklungsprozesses durchzuführen. Aber auch in diesem Fall ist es möglich, Rasch-Analysen durchzuführen, indem man den Erprobungen Ankeraufgaben beifügt, die der letzten durchgeführten Prüfung entnommen wurden.

iii) Interpretation der Scores

Bei der Rasch-Analyse wird die Interpretation des Gesamtpunktwertes (Rohwerte) mithilfe einer Transformationstabelle vorgenommen, die das Analyseprogramm erstellt und die die Rohwerte in Fähigkeitswerte umrechnet (im Analyseprogramm Bigsteps, der Vorversion von Winsteps, ist dies die Tabelle 22 der Analysedatei). Der Nutzen dieser Tabelle liegt darin, dass sie das Problem löst, entscheiden zu müssen, wie die Gesamtpunktwerte aus verschiedenen Testversionen zu interpretieren sind. Jeder Punktwert in einem Test erfordert einen bestimmten Fähigkeitswert, um diesen Punktwert erreichen zu können; je höher die Summe der Punkte, desto höher muss die geforderte Fähigkeit sein, um ihn zu erreichen. Sobald die Items aus allen Tests auf einer gemeinsamen Skala abgebildet wurden, muss man nur den Fähigkeitswert identifizieren, der von den Kandidaten bei einem bestimmten Item gefordert wird, z. B. 60/100. Dann muss man sich die gleiche Tabelle für die anderen Versionen des Tests ansehen und den vorher identifizierten Fähigkeitswert finden. Der von den Kandidaten mit diesem Fähigkeitswert erreichte Punktwert wird dann der Punktwert sein, der 60/100 im ersten Test entspricht. Die Abbildung 9 auf der nächsten Seite ist ein Beispiel für eine Transformationstabelle zur Umrechnung von Rohwerten (linke Zahl in der Spalte) in Fähigkeitswerte (mittlere Zahl in der Spalte) für einen Test mit 48 Items. Sie zeigt, dass, wenn man bei diesem Test die Bestehensgrenze bei einem Punktwert von 30 festsetzen würde, die Bestehensgrenze bei allen weiteren Versionen des Tests bei demjenigen Punktwert gesetzt werden muss, der einem Fähigkeitswert von .71 entspricht. Dies gilt unter der Voraussetzung, dass Ankeraufgaben verwendet wurden, um die Items aus allen Tests auf einer gemeinsamen Schwierigkeitsskala abzubilden, indem man das weiter oben beschriebene Verfahren anwendet.

[2] Anmerkung (TestDaF-Institut): betrifft das Programm Winsteps von Linacre.

120 PERSONS 48 ITEMS ANALYSED: 120 PERSONS 48 ITEMS 2 CATEGORIES

TABLE OF MEASURES ON COMPLETE TEST

SCORE	Measure	S.E.	SCORE	Measure	S.E.	SCORE	Measure	S.E.
0	-5.45E	1.43	17	-.82	.35	34	1.21	.36
1	-4.72	1.03	18	-.70	.35	35	1.35	.37
2	-3.95	.75	19	-.58	.35	36	1.49	.38
3	-3.48	.63	20	-.46	.34	37	1.64	.39
4	-3.12	.56	21	-.34	.34	38	1.80	.40
5	-2.83	.52	22	-.23	.34	39	1.97	.42
6	-2.58	.48	23	-.11	.34	40	2.15	.43
7	-2.36	.45	24	.01	.34	41	2.35	.45
8	-2.16	.43	25	.12	.34	42	2.57	.48
9	-1.97	.42	26	.24	.34	43	2.82	.52
10	-1.80	.40	27	.35	.34	44	3.11	.56
11	-1.65	.39	28	.47	.34	45	3.47	.64
12	-1.49	.38	29	.59	.34	46	3.95	.76
13	-1.35	.37	30	.71	.35	47	4.72	1.03
14	-1.21	.37	31	.83	.35	48	5.45E	1.43
15	-1.08	.36	32	.95	.35			
16	-.95	.36	33	1.08	.36			

Abbildung 9: Transformationstabelle zur Umrechnung von Rohwerten in Fähigkeitswerte

Wenn neue Items zusammen mit Ankeraufgaben erprobt wurden, können sie aus der Erprobungsfassung herausgenommen und in einer Itembank gespeichert werden. Durch den Aufbau einer gut bestückten Itembank wird das Potenzial der Testerstellung erheblich erhöht.

f) Itembank

Eine Itembank ist wie eine Datenbank. Sie ermöglicht es, Items in organisierter Form aufzubewahren und erlaubt eine schnelle und einfache Auswahl von Items, die der Zielbeschreibung einer bestimmten Prüfung entsprechen.

Früher wurden Testmaterialien nach traditioneller Methode für eine bestimmte Prüfung produziert und aufbewahrt. Der Vorteil der Itembank demgegenüber liegt darin, dass Items zusammen aufbewahrt werden, die für viele Tests von unterschiedlicher Schwierigkeit in Auftrag gegeben wurden. Dies erlaubt es, ohne Probleme Items für eine neue Prüfung zusammenzustellen, die z. B. im Schwierigkeitsgrad zwischen zwei existierenden Prüfungen liegt, oder eine größere Schwierigkeitsspanne als die bisherigen Prüfungen hat.

Die Möglichkeiten einer Itembank betreffen aber nicht nur die Effizienz der Testerstellung. Sie erlauben auch eine *qualitative* Verbesserung gegenüber herkömmlichen Verfahren der Testerstellung, weil alle Items in der Bank kalibriert sind. Das heißt, dass die Schwierigkeitswerte aller Items mit Bezug auf eine gemeinsame Schwierigkeitsskala ausgewiesen werden. Die Leistung eines Kandidaten in einem Test, der aus der Itembank zusammengestellt wurde, kann als Position auf dieser gemeinsamen Skala ausgewiesen werden, die gleichzeitig die Itemschwierigkeit und die Personenfähigkeit definiert. Dadurch, dass die Itembank uns ermöglicht, die Lerner präzise auf einer gemeinsamen Fähigkeitsskala abzubilden, was wir subjektiv ohnehin tun, indem wir Bezeichnungen wie „Anfänger“ und „Mittelstufe“ verwenden, leistet sie einen erheblichen Beitrag zur Lösung des Problems, wie Leistungen in unterschiedlichen Tests zu interpretieren sind. Mit anderen Worten, sie bietet eine praktische Lösung für das Problem der Äquivalenz verschiedener Versionen einer Prüfung.

Außerdem kann eine Itembank uns auch Auskunft über die Beschaffenheit einer getesteten Fähigkeit geben. Da die Items nach Schwierigkeiten angeordnet sind, erlauben sie eine detaillierte Beschreibung der getesteten Merkmale (in Form von Items, die Lerner typischerweise auf verschiedenen Stufen lösen können). Deshalb wird eine Itembank auch zu Forschungszwecken und bei Studien zur Konstruktvalidität eingesetzt.

i) Itembank und Testzusammenstellung

Um bei der Zusammenstellung von Testaufgaben für eine Testdurchführung das Itembankverfahren anwenden zu können, müssen folgende drei Voraussetzungen erfüllt sein:

- es muss ein System geben, um Informationen über das Funktionieren der Items zu sammeln, damit sie kalibriert werden können
- statistische Verfahren müssen angewendet werden, um die Itemschwierigkeiten zu ermitteln und sie auf einer gemeinsamen Skala zu verankern
- es muss ein System geben, um die Informationen zu den Items zu speichern und aufzurufen

Die erste Voraussetzung wird durch die Erprobungen erfüllt und die zweite durch die Rasch-Analyse der Erprobungsdaten. Eine Itembank-Software ermöglicht es, Items zu speichern und aufzurufen.

ii) Beispiel für eine Itembank-Software

Das folgende Itembanksystem (IBS) kann als typisches Beispiel dafür gelten, was eine Itembank-Software bieten muss. Es weist die folgenden Hauptmerkmale auf:

- es läuft mit einem aktuellen Windows-Programm
- es kann flexibel mit einer Reihe von Itemtypen umgehen: mit jedem Typ von Einzelaufgabe und auch mit Gruppen von Items zu einem Text
- der Benutzer kann Attribute definieren, um Items zu beschreiben und so ein Beschreibungssystem entwickeln, das so komplex wie nötig ist
- ein voll entwickeltes Suchprogramm erlaubt einen effizienten Zugriff auf die Items
- es erstellt eine Reihe von nützlichen Berichten, u. a. eine Tabelle, in der die in einem Test erreichten Gesamtpunktwerte den geschätzten Rasch-Fähigkeitswerten gleichgesetzt werden
- es speichert, wann und wo die Items verwendet wurden

Itemtypen in der Bank

Das IBS kann drei Gattungen von Itemtypen unterscheiden. Die Abbildung 10 zeigt diese drei Gattungen.

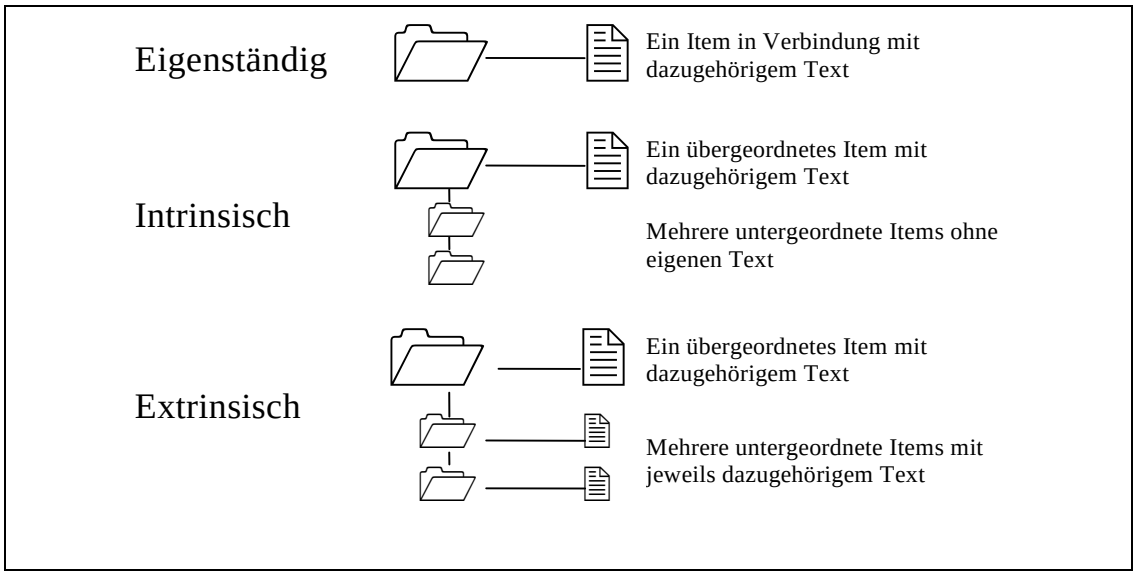

Abbildung 10: Drei formale Itemtypen erfassen alle möglichen Situationen

Ein Beispiel für ein intrinsisches Item wäre ein Cloze-Test. Jede Lücke im Text ist ein untergeordnetes Item, das zusammen mit seinem Schwierigkeitswert einzeln in die Bank aufgenommen wird.

Ein Beispiel für ein extrinsisches Item wäre ein schriftlicher Text, zu dem Aufgaben gestellt werden. In diesem Fall stehen die Aufgaben zwar in Verbindung mit dem Text, könnten aber theoretisch von ihm getrennt werden.

Das System der Attribute

Attribute sind die Merkmale, die das Material in der Bank beschreiben. Sie können sich auf Items, Textpassagen, Grafiken oder jedes andere Merkmal des Materials beziehen, das festgehalten werden soll. Attribute können Nummern oder Wörter sein. Man kann auch die Validierung durch Attribute kennzeichnen. Zum Beispiel könnte das Attribut TRENNSCHÄRFE als eine Dezimalzahl mit einem Wert zwischen -1 und +1 gekennzeichnet werden. Attribute für Wörter können sich auf eine Liste beziehen oder frei vergeben werden, z. B. könnte sich ein Attribut wie THEMA auf eine Liste beziehen, die Themenbereiche, wie z. B. Gesundheit, Freizeit, Wissenschaft, Politik, aufzählt.

Das Suchsystem

Das Herzstück jeder Itembank ist das Suchsystem. Es ermöglicht die Auswahl von Materialien und die Darstellung auf dem Bildschirm zum Zwecke der Testerstellung oder auch zu Forschungszwecken. Es ist wichtig, dass das Suchsystem so umfassend wie möglich ist. Suchaufforderungen könnten folgendermaßen aufgebaut sein: WÄHLE 10 ITEMS MIT SCHWIERIGKEIT ZWISCHEN 50 UND 60 UND THEMA GESUNDHEIT.

Es gibt praktisch keine Grenzen für die Komplexität von Suchaufforderungen. Sie können gespeichert und wieder abgerufen werden. Außerdem können aufeinander folgende Suchaufforderungen eingegeben werden, um die entsprechenden Items einer Liste hinzuzufügen.

Gibt es mehr Items in der Bank als durch die Suchaufforderung spezifiziert werden, werden die Items willkürlich ausgewählt. Der Benutzer kann dann die Liste aller zur Suchaufforderung passenden Items und die Liste der ausgewählten Items ansehen und Items zwischen diesen beiden Listen verschieben.

Berichte

Nachdem ein Test zusammengestellt wurde, können verschiedene Berichte erstellt werden. Zwei der wichtigsten sind der Bericht über die **Fähigkeitswerte** (vgl. Abbildung 11) und über die **Verteilung** (letzteres betrifft die in einer Prüfung erreichten Endnoten, vgl. Abbildung 12).

Note	Punktwert	Fähigkeit	Standardmessfehler
	29	97.82	9.32
A	28	91.27	6.76
B	27	87.25	5.65
	26	84.25	5.02
	25	81.81	4.60
C	24	79.72	4.30
	23	77.86	4.08
	22	76.18	3.91
	21	74.62	3.77
	20	73.15	3.67
D	19	71.75	3.59
	18	70.40	3.53
	17	69.09	3.49
	16	67.81	3.47
	15	66.54	3.45

Abbildung 11: Auszug aus einem Bericht über die Fähigkeitswerte für einen Test mit 30 Items

Die Spalte mit den Fähigkeitswerten entstammt den statistischen Analysen, die nach den Erprobungen durchgeführt und dann in der Bank gespeichert wurden; die Spalte für den Standardmessfehler stellt die Höhe des Messfehlers an bestimmten Punkten der Skala dar. Man kann die Schnittstellen oder Bestehensgrenzen zwischen den Endnoten mithilfe der Fähigkeitswerte festlegen und dann, nachdem man den Test zusammengestellt hat, auf einen Blick sehen, welchen Punktwert von 30 möglichen Punkten der Kandidat erreichen muss, um eine bestimmte Fähigkeit nachzuweisen und eine bestimmte Endnote zu erreichen. Dies bedeutet, dass die Schnittstellen zwischen den Noten festgelegt wurden, bevor der Test durchgeführt wurde, was wiederum präzise und zuverlässige Erprobungsdaten voraussetzt.

Der Bericht über die **Verteilung** gibt einen guten Überblick über die Schwierigkeitswerte der Items im Test. Abbildung 12 auf der nächsten Seite zeigt einen Auszug aus diesem Bericht für den gleichen Test mit 30 Items.

TEST: FCE 1 (ENDNOTEN)

```
A  | 96 |
-------------------------------------------------------
B  | 90 |
-------------------------------------------------------
C  | 80 |
   | 79 |    i260
   | 78 |    i259     i258
   | 77 |    i253
   | 75 |    i250
   | 74 |    i242
   | 73 |    i237     i240
-------------------------------------------------------
D  | 72 |    i236
   | 71 |    i227     i233
   | 69 |    i220
   | 68 |    i218     i216
   | 67 |    i209     i210
   | 66 |    i203
```

Abbildung 12: Auszug aus einem Bericht zur Verteilung

Der Bericht über die Verteilung ordnet die Items nach Schwierigkeit und bildet sie auf der gemeinsamen Skala ab, die auch dem Bericht über die Fähigkeitswerte (Abbildung 11) zugrunde liegt. Der Bericht zeigt, wie sich die Items in einem Test verteilen (die Buchstaben in der ersten Spalte beziehen sich auf die Endnoten, die man im *First Certificate in English* erreichen kann).

iii) Testzusammenstellung ohne Itembank

Das oben Gesagte verdeutlicht, wie sehr eine computerbasierte Itembank den Testentwicklern bei der Zusammenstellung der Tests hilft.

Wenn keine Itembank zur Verfügung steht, wird der Testentwickler vermutlich eine Redaktionssitzung einberufen, in der der Test zusammengestellt wird. Daran nehmen in der Regel zwei bis drei Personen teil, die die Prüfung sehr gut kennen, und möglicherweise ein externer Berater für die Prüfung bzw. für den entsprechenden Prüfungsteil.

Die Prüfung bzw. der Prüfungsteil wird entsprechend der Testbeschreibung mit den bearbeiteten und erprobten Testaufgaben aus früheren Arbeitsaufträgen zusammengestellt. Dabei muss nicht nur die Breite der Themen in der gesamten Prüfung berücksichtigt werden, sondern auch bereits verwendete Themen in zurückliegenden Prüfungen, damit die Auswahl geeigneter Themen genügend Abwechslung bietet. Sofern in einem Prüfungsteil Einzelitems vorkommen, z. B. 20 Multiple-Choice-Items, die verschiedene grammatische Strukturen prüfen, sollte man auf eine gewisse Bandbreite von Strukturen und unterschiedliche Schwierigkeitswerte achten. Wenn die Prüfung von vielen Kandidaten aus unterschiedlichen Ländern abgelegt wird und die Kandidaten unterschiedlichen Altersgruppen angehören, sollte die Auswahl der Themen und der Items nicht eine bestimmte Gruppe bevorzugen, obwohl natürlich immer einige Kandidaten bestimmte Texte interessanter finden als andere.

Zum Schluss werden die ausgewählten Texte und Items in das Layout gebracht, das für die Prüfung festgelegt wurde, und der Test geht anschließend in Druck. Die gedruckten Testexemplare müssen bis zum Prüfungstermin sicher aufbewahrt werden.

4. ÜBUNGEN

1. Es soll eine Beschreibung für einen Test zum Sprechen auf fortgeschrittenem Niveau (ALTE 4 oder 5 oder C1 bzw. C2) entwickelt werden. Der Test ist als Sprachnachweis für ausländische Studierende zur Aufnahme an einer Hochschule gedacht.

   Schauen Sie auf die fachlichen Überlegungen auf der nächsten Seite unter 1 (Lösungsvorschläge für die Übungen).

   Welche Auswirkung wird Ihrer Meinung nach jede dieser Überlegungen auf die Entscheidungen haben, die Sie in Bezug auf den Test fällen müssen?

2. Es ist ein Sprachtest auf mittlerem Niveau (ALTE 3 oder B2) geplant, der sich vorwiegend an Kandidaten wendet, die zwischen 15 und 20 Jahre alt sind und unterschiedliche Staatsangehörigkeiten besitzen.

   Welche Themenbereiche würden Sie für angemessen halten, um sie in Texten und Aufgaben zum Lese- und Hörverstehen zu verwenden?

   Welche Textquellen könnte man den Testautoren vorschlagen?

3. Die Texte in Anhang A wurden alle in der Phase der Begutachtung abgelehnt. Sie wurden für einen Subtest zum Leseverstehen in einer Feststellungsprüfung eingereicht, die sich an jüngere Erwachsene wendet (die vermutlich eher berufstätig sind, als dass sie studieren). Die Kandidaten besitzen unterschiedliche Staatsangehörigkeiten und können die Prüfung sowohl im Heimatland als auch im Land der Zielsprache ablegen.

4. Texte, die für eine Prüfung ausgewählt werden, sind häufig leicht bearbeitet oder vereinfacht, aber sie sollten nicht stark umgeschrieben werden. Die Texte haben normalerweise eine Länge von 250-300 Wörtern und die Testautoren werden beauftragt, fünf Multiple-Choice-Items zu jedem Text zu schreiben.

   Welche Gründe würden Sie für die Entscheidung nennen, jeden dieser Texte abzulehnen?

5. In Anhang B finden Sie einen Ausdruck aus dem Iteman-Analyseprogramm (MicroCAT). Er zeigt, wie einige Multiple-Choice-Items bei einer Erprobung funktioniert haben. Geben Sie, soweit Sie können, Auskunft über jedes Item, indem Sie die folgenden Fragen beantworten:

   a) **Wie schwierig war das Item für die Kandidaten?**
   b) **Wie trennscharf war es?**
   c) **Bildeten die richtige Antwort und die Distraktoren eine gute Verteilung der Optionen?**

Lösungsvorschläge für die Übungen

1.

Die erste fachliche Überlegung betrifft die Art der Realsituationen, in denen die Kandidaten zukünftig die Fremdsprache gebrauchen müssen. Im genannten Fall werden sie über akademische Angelegenheiten mit Angestellten der Universität oder ihren Mitstudenten sprechen müssen, wobei es darum geht, Meinungen auszudrücken und zu rechtfertigen, anderen zuzustimmen und Meinungsverschiedenheiten mit einem angemessenen Grad an Höflichkeit auszudrücken. Sie werden sich auch in Situationen befinden, in denen es um das tägliche Zurechtkommen in einem fremden Land geht (eine Unterkunft suchen, einkaufen usw.) und um die Kontaktpflege mit anderen Studenten.

Die zweite fachliche Überlegung betrifft das sprachliche Niveau, das die oben genannten Situationen erfordern. Am wichtigsten ist hier die Fähigkeit, die Anforderungen des Studiums zu meistern. Wenn Nicht-Muttersprachler an Diskussionen mit Muttersprachlern über akademische Themen teilnehmen müssen, erfordert dies ein fortgeschrittenes sprachliches Niveau. Für die Alltagssituationen reicht vielleicht ein niedrigeres Niveau.

Die dritte fachliche Überlegung betrifft die realen kommunikativen Aktivitäten, die im Test abgebildet werden müssen. Bei dieser Prüfung ist es relevant, Situationen abzubilden, die folgende Sprechhandlungen betreffen: begründen, diskutieren, Meinungen äußern, zustimmen bzw. sagen, dass man anderer Meinung ist, sowie nach Informationen fragen bzw. sie geben. Der Test könnte eine Partnerprüfung sein, damit man Aufgaben entwickeln kann, die die Kandidaten in Diskussionen verwickeln und die erfordern, dass ein Kandidat über die Meinung des anderen berichtet.

Die vierte fachliche Überlegung betrifft die Informationen, die den Testanwendern zur Verfügung gestellt werden müssen. Ob eine Universität, die diesen Test einsetzt, den Kandidaten im Falle eines Nicht-Bestehens eine ausführliche Rückmeldung über die Gründe für ihr Scheitern gibt oder nicht, hat etwas damit zu tun, ob man eine spezielle Vorbereitung auf diese Prüfung für notwendig erachtet oder nicht.

2.

Mögliche Themenbereiche sind:

Familie, Freunde, zu Hause in der Familie leben, Hobbys und Freizeitinteressen, Sport und Spiele, Lernen und Schule, berufliche Laufbahn und zukünftige Ziele, Unterhaltung und kulturelle Interessen (z. B. Musik, Theater, Kino, Lesen), Ferien, Reisen, Feste und Feiern, Kleidung und Mode, Einkaufen, Nahrungsmittel und Kochen, Gebäude und Orte, einige Themen zu Umweltschutz und sozialen Problemen (z. B. im Hinblick auf den Tourismus), Leben in der Stadt und auf dem Land

Mögliche Textquellen sind:

Zeitschriften, Broschüren, Zeitungen, Werbematerial, Literatur und Sachbücher, Programme im Radio

3.

Die folgenden Gründe zur Ablehnung von Texten sollten berücksichtigt werden:

Text 1: Das Thema ist nicht angemessen und nur für Nordeuropäer von wirklichem Interesse. Es könnte auch diejenigen Kandidaten aufregen oder belasten, die Krebs haben (oder die einen krebskranken Verwandten haben).

Text 2: Texte, die sich aus vielen kleinen Einzeltexten zusammensetzen, sind im Prinzip verwendbar. In diesem Fall nehmen jedoch die Adressen sehr viel Raum ein und die übrigen Informationen geben nicht genügend her, um Items zu entwickeln, die über ein sehr niedriges Niveau hinausgehen. Der Stil ist witzig, sehr journalistisch und mit vielen Wortspielen, sodass er möglicherweise für die Lerner nicht verständlich ist, z. B. „kiss-me-quick-Newquay", „an unbeatable natural high" usw. Witzelnde journalistische Texte sind aus diesen Gründen oft unangemessen, obwohl sie auf den ersten Blick attraktiv erscheinen.

Text 3: Das Thema ist für Kandidaten aus unterschiedlichen Kulturen nicht angemessen; viele sind nicht vertraut damit, dass Tiere geliebte Haustiere sind, und betrachten besonders Hunde als schmutzig. Weitere Probleme sind kulturspezifische Gesellschaften wie die RSPCA und der Brauch, Tiere als Weihnachtsgeschenk zu verschenken.

Der Artikel scheint sich auch eher an Kinder und Eltern zu wenden als an jüngere Erwachsene, die zwischen diesen beiden Altersgruppen liegen.

Wörter wie „romping", „cute, pudgy puppy", „floppy ears", „doggy best friend" müssten umformuliert werden.

Text 4: „Auswärts essen gehen" könnte ein geeignetes Thema für diese Zielgruppe sein, aber der Artikel ist sehr kulturspezifisch. Die Vorstellung, dass es interessant ist, in Restaurants zu essen, die Gerichte aus verschiedenen Ländern anbieten, ist nicht universell verbreitet und dort, wo dies üblich ist, vermutlich spezifisch für bestimmte Gesellschaftsschichten. Bei den Lesern von Zeitschriften wie „Time Out" wird viel gemeinsames kulturelles Hintergrundwissen vorausgesetzt, wie die Aussage „committed but not worthy" über ein vegetarisches Restaurant zeigt. Außerdem ist es besser, Alkohol nicht zu erwähnen, wenn Kandidaten aus Ländern den Test ablegen, in denen Alkohol verboten ist.

Es gibt sehr viele Fremdwörter wie „sushi", „sashimi", „calzone", die einige Kandidaten bevorzugen und andere benachteiligen. Umgangssprachliche Ausdrücke wie „pit-stop", „booze" und „shoestring" sind Kennzeichen eines journalistischen Stils und deshalb problematisch.

Text 5: Das Thema Sport scheint für diese Zielgruppe geeignet zu sein, aber der Artikel ist so spezialisiert, dass er unverständlich wird. Er verwendet auch sehr viele umgangssprachliche Ausdrücke, die eher der gesprochenen als der geschriebenen Sprache entstammen, und der humorvolle Ton wird vermutlich die Probleme beim Verstehen noch verschärfen.

Text 6: Themen wie die Tier- und Pflanzenwelt, die Erforschung der Vergangenheit und andere populärwissenschaftliche Themen eignen sich oft als Prüfungstexte. Das gewählte Thema ist unkontrovers und von allgemeinem Interesse, aber kann man mit diesem Text eine ausreichende Anzahl an Multiple-Choice-Items entwickeln? Der Text sagt dem Leser eigentlich nicht viel mehr, als dass Fußspuren von einem Dinosaurier gefunden wurden. Der Inhalt ist ziemlich dürftig und es gibt eine Reihe schwieriger Wörter wie „dubbed", „fared" und „clams".

4.

Anmerkungen zu den MicroCAT-Beispielen:

Item 1: Dies war ein leichtes Item (proportion correct = .92). Es war nicht besonders trennscharf (point biserial correlation = .25, d. h. gerade noch akzeptabel). Die Verteilung der Kandidatenantworten zwischen der richtigen Lösung und den Distraktoren war nicht gut, da die meisten die richtige Lösung gewählt haben; der Distraktor C wurde von niemandem gewählt und der Distraktor B nur von .01 Kandidaten.

Item 2: Dies war ein relativ schweres Item (proportion correct = .47). Es war trennscharf (point biserial correlation = .47). Wenn man die Distraktoren zusammen betrachtet, haben mehr Kandidaten die Distraktoren B und C gewählt als die richtige Lösung A.

Item 3: Dies war ein verhältnismäßig leichtes Item (proportion correct = .77), das nicht sehr trennscharf war (point biserial correlation = .29). Einer der Distraktoren (D) wurde von keinem Kandidaten gewählt, was zeigt, dass er ein schwacher und wenig plausibler Distraktor ist.

Item 4: Dies war ein relativ schweres Item (proportion correct = .43), das nicht sehr trennscharf war (point biserial correlation = .12, d. h. nicht akzeptabel, weil zu niedrig). Fast ebenso viele Kandidaten wählten den Distraktor C (.40) wie die richtige Lösung (.43), während sich Distraktor B als ziemlich schwach herausgestellt hat (.01).

Item 5: Dies war ein schweres Item (proportion correct = .38) mit einer guten Trennschärfe (point biserial correlation = .54). Mehr Kandidaten haben den Distraktor A (.41) als die Lösung (.38) gewählt, einige wählten C oder D. Es sieht jedoch so aus, als hätten sich die schwächeren Kandidaten von den Distraktoren verleiten lassen, von daher scheint das Item gut zu funktionieren.

Item 6: Dies war ein ziemlich leichtes Item (proportion correct = .83), auch vergleichsweise trennscharf (point biserial correlation = .37). Jeder Distraktor wurde von einigen Kandidaten gewählt.

ANHANG A

Abgelehnte Texte

1 The Dangerous Sun

We have only recently become more conscious of the damaging effects of the sun. Experts believe that changes in the environment have made exposure to the sun more dangerous. Strong links have been found between thinning of the ozone layer and skin cancer. This thinning will mean that the screening out of the harmful UV radiation will be less successful. This is most alarming, since for every 1 percent increase in UV the incidence of skin cancer goes up by 2 percent.

The rate of skin cancer is increasing rapidly. Sunbathing, for example, is now seen as more hazardous than previously recognised. Health experts believe that increasing exposure to the sun without protection is the major factor in the increase. It has become increasingly urgent to educate the public about the risks of too much exposure to the sun.

Most skin cancers are curable. Those which are not are mostly in one of the three categories of skin cancers - malignant melanoma. Everyone should learn the early signs of skin cancer and if they are discovered a doctor should be consulted. Even with the most serious forms of cancer if diagnosed and treated early there is a very good chance of achieving a cure.

Children's skin is especially delicate and experts believe that long, unprotected exposure to the sun in childhood dramatically increases the risk of skin cancer. Since fifty percent of our lifetime exposure to UV radiation occurs in the first 20 years of life, high factor sun screens become extremely important.

The Australians are very aware of sun-related problems. They have the highest rate of skin cancer in the world and are very advanced in promoting sun awareness. In the UK, the medical profession and manufacturers of beauty and sun care products are helping in a campaign to educate the public about sensible sun protection.

2 A to Z of Activities

Hang Gliding
Hang-gliding courses provide an unbeatable natural high but young people must be over the age of 15 to take part.
Contact: The British Hang Gliding and Paragliding Association (BHPA) c/o Jennifer Burdett, Old School Room, Loughborough Road, Leicester LE4 5PJ. Tel: 0533 611322 or Fax: 0533 611323.

Initiative

Seize it!

Jousting
At a medieval castle near you - being a spectator rather than a participant might be advisable here!
Contact: the British Jousting Centre, c/o Max Diamond, Tapely Park, Instow, North Devon EX39 4NT. Tel: 0271 861200.

Karting
Go, go, go, Karting. Try life in the fast lane and remember this is how Mansell, Prost, Senna et al started their careers.
Contact: The RAC Motor Sports Association Ltd., Motor Sports House, Riverside Park, Colbrook, Slough SL3 0HG. Tel: 0753 681736.

Lifesaving
The Royal Life-Saving Society UK is the country's principal authority on lifeguard training.
Contact: RLSS UK, Mountbatten House, Studley, Warwickshire B80 7NN. Tel: 052.785.3945 or Fax: 052.785.4453.

Mountaineering
From Ben Nevis to Snowdon the challenge is steep.
Contact: UK Mountain Training Board, Capel Curig, Gywynedd, Wales. Tel: 06904.272.

National Trust
With over 600,000 acres of land and 535 miles of coastline, Britain's National Trust is the largest conservation charity in the world. The Education Group Membership provides resources and teaching as well as free access to many Trust properties.
Contact (for a free education pack): The National Trust, 36 Queen Anne's Gate, London SW11 9AS. Tel: 071.222.9251 or Fax: 071.222.5097.

Orienteering
Take a compass bearing and step this way for details of local clubs, activities and a list of publications.
Contact: British Orienteering Federation, Riversdale, Dale Road, North Darley Dale, Matlock, Derbyshire DE4 2HX. Tel: 0629. 734042.

Parachuting
Don't come down to earth with too much of a bump. Minimum age 16.
Contact: British Parachute Association, Kimberley House, 47 Vaughan Way, Leicester LE1 4SG. Tel: 0533.519778.

Quasar
If all else fails, you could try spending some time in a darkened room.
Contact: See Yellow Pages.

Riding
Gallop along to your nearest approved riding school but don't get too saddle sore.
Contact: The British Horse Society, British Equestrian Centre, Stoneleigh, Kenilworth, Warwickshire CV8 2LR. Tel: 0203. 696697 or Fax: 0203 692351.

Disabled young people needn't miss out on the pleasures of horse riding.
Contact: Riding for the Disabled, Avenue R. National Agricultural Centre, Kenilworth, Warwickshire CV8 2LY.

Surfing
The surf's not only up at kiss-me-quick Newquay; how about the Yorkshire coast, South Wales, Scotland and Eire? The British Surfing Association has details of approved surfing courses and holidays designed for young people.
Contact: BSA, c/o John Briant on Tel: 0637. 850737.

Tennis
Could a future Wimbledon champion be lurking in your youth group? There's only one way to find out, and remember, England expects...
Contact: The Lawn Tennis Association, The Queen's Club, West Kensington, London W14 9EG. Tel: 071.385.2366.

Universities
Can provide excellent - and inexpensive - accommodation for non-student groups during the summer holidays.
Contact: British Universities Accommodation Consortium Ltd., Box 1009, University Park, Nottingham NG7 2RD. Tel: 0602.504571 or Fax: 0602.422505.

3 A Dog's Life for You!

Lots of people want a dog of their own. They imagine romping around with a cute, pudgy puppy with big brown eyes and floppy ears. Or they dream of racing across sunny hills with their doggy best friend. But owning a dog is an enormous responsibility. The puppy stage only lasts a few weeks, and then you've got a grown creature whose health and happiness depend totally on you and your family.

Dogs need loads of attention, love, and patience. They're expensive, and they also take up a lot of time. The RSPCA says: If you want a dog, think again. Then think once more. After all that thinking, maybe you'll be sure that you could give a dog a good home. Caring for a dog can be very rewarding. But once you've got it, you've got it for its whole lifetime. This is a decision you can't go back on. Getting rid of a dog because you're bored with it, or find it too much work, is horribly unfair.

You've probably never had to think so far head. School days don't finish until you are 16 or 18, and that seems far enough as it is. But if you want a dog, you have to think 15 years on. Because that's how long some dogs can live. So, if you're 10 now, it could still be alive when you're 25!

The chances are that you'll have left home by then. "Parents usually end up looking after their children's pets when their children have gone," says Terence Bate, the RSPCA's chief veterinary officer. His main rule is that everyone should understand exactly what getting a dog means. "Above all else the parents must accept ultimate responsibility for the animal," he says. "That's the most important thing. Even though it's the child that gets a dog, even though it's the child that looks after it and does the day-to-day work."

Terence's second rule is this: never ever give a pet as a surprise for someone. People often say how much they'd love a puppy. Often it's the IDEA of having one that they love. If you told them about all the hard work, they probably wouldn't be so keen. All the attention and excitement wears off after a few weeks, when the puppy is no longer a cute, bouncy novelty. So the dog gets abandoned. The problem is so big that the RSPCA doesn't let anyone adopt a dog or any other animal for a couple of weeks before Christmas, to stop them being given as presents.

4 'Time Out' Eating and Drinking Awards 1994

We're getting closer to dishing out the *Time Out* Eating and Drinking Awards.

This week we're looking at places to eat on a budget - always one of the most popular categories of the awards. The following are some of the best places to find a good meal for around £12-a-head. Which one gets your vote?

Best Budget Meal Award

Eco *162 Clapham High Street, SW4 (071 978 1108) Clapham Common tube. (Open Mon-Thur 11.30am - 3pm, 6.30-11pm: Fri, Sat 11.30am -4pm, 6.30-11.30pm: Sun 12noon-4.30pm 6.30-10.30pm*).
This pizzeria in Clapham achieved instant success when it opened last year - so much so that the stunningly designed interior has already been extended. Pizzas are outstandingly good and start at £3.40 for a margherita, up to £5.90 for a seafood calzone. House wine is £6.50. High style for shoestring prices.

Harvey's Cafe: *358 Fulham Road, SW10 (071 3520625), Fulham Broadway tube (Open Tue-St 12.30 -5pm, 7.30-11pm: Sun 12noon-3.30pm).*
A short and very reasonably priced menu of fashionably rustic cooking (chick pea and sun-dried tomato soup, risottos, home-made ice-creams) is served in this light, bright and friendly, first-floor restaurant. A la carte dinner is around £15 with wine (£6.95 for house). Lunch is even more of a snip - for £5.95, you get bread, olives and two courses.

Heather's Cafe-Bistro *190 Trundley's Road, SE8 (081 691 6665) Surrey Quays tube or New Cross tube/BR/225 bus. (Open Wed-Sat 7-10.30pm: Sun 12.30-6pm).*
Unlicensed vegan and vegetarian restaurant that puts on a spread to defy any sceptical carnivore. It's an all-in price of £7.50 a head (£4 for children) for soup, a selection of good-looking starters, and thoughtfully-prepared and original main courses. If you want to drink, take your own booze and pay 30p corkage. Committed but not worthy, the restaurant is brightly-decorated and candlelit.

Tokyo Diner *2 Newport Place, WC2 (071 287 8777) Leicester Square tube (Open daily 12noon - 12midnight).*
A welcome West End pit-stop. This ground floor and basement Japanese cafe is commendably good value. The menu's carefully annotated and covers a wide range of popular dishes including bento boxes (£6.90-£11.90), sushi (£4.50/£6) and sashimi (£7). It also has some of the cheapest Japanese beers in town - they're all £1.90. Tips are neither expected or accepted - if only more restaurants followed suit!

5 The Worst Damn Job in Baseball

God I'd hate to be an umpire. I was once, for 3 innings, and my team-mates wouldn't speak to me afterwards, *writes Dennis O'Neill*.

Let's face it, umpiring is the hardest job in baseball, bar none. You're the only one who has to be out there for the whole nine innings, regardless of who's batting. You even have to be awake *between* innings, to check that the pitcher doesn't take too many warm-ups, to jolly the game along a little, whatever. And you don't get to take off that hot, heavy gear, either. Even catchers get a break when they go up to bat.

It's a dangerous job, too. Rule 1.16(d) states that all catchers must wear a protective helmet when fielding their position. But an umpire? Hey, if he gets hit in the head, at least it won't hurt him. Might lose some sawdust, no big deal. And if anyone argues, you can always throw him out. Not that it's ever stopped anyone. Ok, no one in Britain can bait an ump with the ferocity that Billy Martin used to employ, nor with the wily cussedness of Earl Weaver stealing third base and refusing to give it back.

Even so, two umps I know in Leeds packed it in because they got fed up with everyone disputing every call every time.

I can see their point. It can't be easy to face a nerd, who couldn't hit a Little League fastball if he was swinging a Giant Sequoia, telling you what the rules are. Some things you just don't need.

I mean, it's hard enough as it is to decide whether or not an 80mph projectile aimed at your face has crossed through an imaginary rectangle about 250 times a game without having your every choice questioned. A typical innings might consist of three or four "You called that a strike?!"; half a dozen or so of "Well where the hell is the strike zone, then?" and "No way, just no way was he safe/out!" at least twice. After an hour or two in the sun (or, as is more likely in Britain, in the teeth of a force nine gale), that can start to wear you down.

6 Footprints in Time

Dinosaurs have left their mark on the Earth

A storm lingered on the horizon as a herd of Apatosaurus - huge plant-eating dinosaurs with long flexible necks - followed the shore line of a lake in what is now Colorado, USA. The largest dinosaur led the animals, the smallest were protected in the centre of the herd. As they walked they crushed freshwater clams underfoot and left a trail of deep footprints in the mud.

The herd moved on, its members died: eventually the whole species became extinct. But, buried by layer on layer of fresh mud, the footprints remained, very slowly turning to stone. Frozen in time, the tracks waited to be uncovered in the 20th century, when they would enable palaeontologists to reconstruct the passage of the dinosaur herd 100 million years after it had happened.

Signs of life

Footprints have been vital to our new understanding of dinosaurs. Whereas bones allow scientists to reconstruct the dinosaurs' physical appearance, footprints offer clues to their behaviour - how fast they ran, whether they lived alone or in groups, how they cared for their young, and how they fared in the desperate survival game as hunters or hunted.

Sometimes, patterns of footprints offer 'snapshots' of dramatic encounters. In Texas, one set of tracks seems to show a single giant plant-eating sauropod being pursued by a pack of carnivorous dinosaurs - the sauropod's broad heavy prints are surrounded by the imprint of lighter three-toed hunters. In Queensland, Australia, large numbers of hypsilophontids, small plant-eating dinosaurs, left a chaotic jumble of footprints as they fled in panic from flesh-eating theropods.

Mass migration

Elsewhere, it is the density of the dinosaur tracks that astonishes, revealing the extraordinary numbers of the creatures that once roamed the planet. There are so many dinosaur footprints on the eastern slopes of the Rocky Mountains, in Colorado and New Mexico, that the area has been dubbed a 'dinosaur motorway'. Geologist Martin Lockley believes the millions of tracks record an annual mass migration of dinosaurs, similar to the great movements of wildebeest across the Serengeti Plain in modern-day Africa.

Reconstructing the life of the dinosaurs will always be a work of the imagination. But their footprints are the closest we can come to the living reality of the dinosaurs' world.

ANHANG B

Iteman-Beispiele (MicroCAT) für Übung 4

Seq. No.	Scale - Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1 - 1	.92	.14	.25	A	.07	.13	.01	-.24	
					B	.01	.03	.01	-.08	
					C	.00	.00	.00	-.00	
					D	.92	.84	.98	.25	*
					Other	.00	.00	.00		
2	1 – 2	.47	.58	.47	A	.47	.19	.76	.47	*
					B	.22	.41	.05	-.36	
					C	.26	.29	.18	-.09	
					D	.05	.11	.00	-.22	
					Other	.00	.00	.00		
3	1 – 3	.77	.28	.29	A	.04	.08	.01	-.16	
					B	.77	.64	.92	-.29	*
					C	.18	.28	.07	-.22	
					D	.00	.01	.00	-.22	
					Other	.00	.00	.00		
4	1 – 4	.43	.13	.12	A	.16	.22	.10	-.15	
					B	.01	.01	.00	-.07	
					C	.40	.40	.40	.00	
					D	.43	.36	.49	.12	*
					Other	.00	.00	.00	-.06	
5	1 – 5	.38	.63	.54	A	.41	.54	.26	-.22	
					B	.38	.20	.59	.33	*
					C	.13	.16	.09	-.10	
					D	.08	.09	.06	-.06	
					Other	.00	.00	.00	-.03	
6	1 – 6	.83	.30	.37	A	.83	.66	.96	.37	*
					B	.06	.12	.01	-.23	
					C	.07	.13	.02	-.17	
					D	.03	.08	.01	-.20	
					Other	.00	.00	.00	-.03	

ANHANG C

Literaturempfehlungen

Bachman, L./Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, D. (1989). *Language Testing, A Critical Survey and Practical Guide*. London: Edward Arnold.

Carroll, B. J./Hall, P. J. (1985). *Make Your Own Language Tests*. Oxford: Pergamon Press.

Henning, G. (1987). *A Guide to Language Testing*. New York: Newbury House.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

ANHANG D

Ein Beispiel für die Testentwicklung
Entwicklung eines Tests für die Fertigkeit Sprechen

Dieser Abschnitt soll die Verfahren beschreiben, die bei der Entwicklung eines Tests zum Sprechen angewendet werden. Viele der angesprochenen Aspekte gelten aber auch für die Testentwicklung allgemein.

In einem ersten Schritt wird der Bedarf für einen neuen Test zum Sprechen festgestellt. Es kann z. B. vorkommen, dass ein existierender Test durch einen neuen ersetzt werden muss. Ist dies der Fall, werden die Testentwickler schon einige klare Vorstellungen darüber haben, welche Verbesserungen der neue Test gegenüber dem alten aufweisen sollte. Vielleicht ist es notwendig, Veränderungen in der Testpopulation zu berücksichtigen oder neuere Entwicklungen in der Testtheorie aufzugreifen. Vielleicht muss man auf Änderungen bei der Prüfungsdurchführung eingehen; es kann z. B. sein, dass für den Test zukünftig weniger Zeit zur Verfügung steht oder dass weniger Prüfer rekrutiert werden können. Gründe für eine Veränderung können theoretischer oder praktischer Natur sein, sie tragen in jedem Fall zur Festlegung des Formats bei.

Auch wenn der Test völlig neu ist, gibt es bekannte Faktoren, die die Entwicklung von Anfang an beeinflussen. Folgendes könnte ein Beispiel dafür sein: Eine Universität hat einen Test zum Schreiben eingesetzt, um die Sprachkenntnisse ihrer ausländischen Bewerber zu testen, die notwendig sind, um ein Studium aufzunehmen. Es wurde aber festgestellt, dass diejenigen, die den Test bestanden haben, dennoch große Schwierigkeiten in Seminaren und Tutorien hatten und dass einige ausländische Studenten aufgrund ihrer schlechten Sprachkenntnisse sozial isoliert waren. Wenn in diesem Fall entschieden wird, dem bisherigen Auswahlverfahren für Bewerber einen Test zum Sprechen hinzuzufügen, haben die Testentwickler schon eine klare Vorstellung über das geforderte sprachliche Niveau, die Fertigkeitsbereiche, die Themen sowie die Situationen, die der Test erfassen muss. Sie haben eine Vorstellung davon, was im Kontext dieses Tests nützlich ist. An diesem Punkt beginnt der Prozess der Testentwicklung.

1. DIE PLANUNGSPHASE

Es wird eine Situationsanalyse durchgeführt, die Folgendes untersucht und beschreibt:

a) Die von dem Test Betroffenen. Dies sind alle, die am Test beteiligt sind, also die Testentwickler, diejenigen, die den Test durchführen, die Kandidaten und diejenigen, die die Ergebnisse verwenden. Dazu gehören Lerner, Lehrer, Eltern, Schuldirektoren, Regierungsbehörden und kommerzielle Unternehmen. Der Test muss von all diesen Personen angenommen werden, d. h., sie müssen verstehen und akzeptieren, warum der Test so ist, wie er ist.

b) Ziel und Zweck des Tests. Es ist wichtig, eine klare Vorstellung darüber zu haben, warum der neue Test zum Sprechen gebraucht wird und warum er ein bestimmtes Format haben soll. Diejenigen, die für seine Entwicklung zuständig sind, müssen seine Merkmale verantworten können. Es sollte festgestellt werden, auf welchem Schwierigkeitsniveau der Test liegt und wie gut er im Hinblick auf das Curriculum und die gebräuchlichen Unterrichtsmethoden in das gegenwärtige System passt. Außerdem müssen evtl. zukünftig geplante Entwicklungen identifiziert werden.

c) Äußere Einflüsse. Es muss überlegt werden, wie das Sprechen in dem Kontext geprüft werden sollte, in dem der Test eingesetzt wird, dabei sollten bereits auf dem Markt verfügbare Tests berücksichtigt werden. Außerdem sollten die bildungspolitischen Anforderungen, die gegenwärtigen sozio-ökonomischen Bedingungen und die Verhältnisse vor Ort bedacht werden.

d) Interne Faktoren. Der neue Test zum Sprechen kann z. B. für eine Schule, eine Universität oder ein Ministerium entwickelt werden. Wer immer der Auftraggeber ist, der Test muss mit den vorherrschenden Arbeitsverfahren der jeweiligen Institution übereinstimmen und die dort

vorhandenen Fachkenntnisse im Testen des mündlichen Ausdrucks berücksichtigen. Auch die Mittel, die dem Testentwickler im Hinblick auf Personal, Technologie, Zeit und Geld für die Testentwicklung, die Durchführung, die Bekanntgabe der Ergebnisse, das Erstellen weiterer äquivalenter Tests und die Validierung zur Verfügung stehen, müssen bedacht werden.

Ein Projektplan muss ausgearbeitet werden, der die Ziele nennt, die notwendigen Mittel spezifiziert und den zeitlichen Ablauf festlegt. Das folgende Beispiel zeigt die Phasen eines Projektplans für die Entwicklung eines Subtests zum Sprechen in einer Prüfung, die aus mehreren fertigkeitsspezifischen Subtests besteht.

Phase 1: Entwickeln der Testbeschreibung für den Subtest Sprechen

Anweisungen zum Test schreiben
Intern verteilen
Revidieren und neu schreiben
Ersten Entwurf der Testbeschreibung schreiben
An Mitglieder des Projektgremiums verteilen
Testbeschreibung revidieren und in ihre endgültige Form bringen

Materialien für einen Modelltest entwickeln
1. Treffen: Die Testautoren über die benötigten Itemtypen informieren und sie schulen
2. Treffen: Die Testautoren reichen ihre Entwürfe ein
Items bearbeiten
Durchführungsbestimmungen schreiben
Test und Bewertungsunterlagen drucken

Den Modelltest erproben
Mögliche Erprobungszentren kontaktieren
Ein oder zwei geeignete Zentren in der Nähe auswählen
Mündliche Prüfung durchführen
Ergebnisse analysieren
Aufgaben auswählen und in den Entwurf der Testbeschreibung aufnehmen

Planen der Phase 2: Weitere Erprobungen
Hinweise für Testautoren schreiben
Instruktionen und Trainingsmaterialien für Testautoren schreiben
Team von Testautoren finden
Testautoren instruieren und schulen
Begutachtungsverfahren für die Items schriftlich festlegen
Art der Rückmeldung festlegen, die den Testautoren über das Verhalten der Items gegeben wird
Sitzungen festsetzen, in denen eine intensive Produktion von Items stattfinden soll

Phase 2, Teil 1: Weitere Erprobungen

Die Items erstellen
1. Treffen: Testautoren über die Items informieren, die für Erprobungen benötigt werden, und sie schulen
2. Treffen: Testautoren reichen ihre Entwürfe ein
Items bearbeiten
Durchführungsbestimmungen überarbeiten
Test und Bewertungsunterlagen drucken

Die Items erproben
Mögliche Testzentren kontaktieren
Ein oder zwei geeignete Zentren in der Nähe auswählen
Mündliche Prüfung durchführen
Ergebnisse analysieren
Items für die Itembank auswählen

Phase 2, Teil 2: Weitere Erprobungen

Die Items erstellen
1. Treffen: Testautoren über die Items informieren, die für Erprobungen benötigt werden, und sie schulen
2. Treffen: Testautoren reichen ihre Entwürfe ein
Items bearbeiten
Test drucken

Die Items erproben
Mögliche Testzentren kontaktieren
Ein oder zwei geeignete Zentren in der Nähe auswählen
Mündliche Prüfung durchführen
Ergebnisse analysieren
Items für die Itembank auswählen

Phase 3: Training für Prüfer

Anleitungen zur Beurteilung des mündlichen Ausdrucks für die Prüfer schreiben
Ein Schulungsvideo für die Prüfer produzieren
Prüfer suchen
Prüferteams schulen
Einen Test für die Zertifizierung von Prüfern entwickeln
Rückmeldung der Prüfer über das Training, die Materialien usw. einholen

Phase 4: Überarbeitung der Materialien für Prüfer

Revidierte Fassung der Anleitungen zur Beurteilung des mündlichen Ausdrucks erstellen
Schulungsvideo auf den neusten Stand bringen
Zertifizierungsverfahren auf den neusten Stand bringen
Weitere Prüfer suchen
Schulungen für neue Prüfer festsetzen und durchführen
Ausbilder für Prüfer in den Ländern finden, in denen der Test eingesetzt wird
Trainingsseminare durchführen, um diese Ausbilder zu schulen

2. DIE KONZEPTIONSPHASE

In der Konzeptionsphase wird eine erste Fassung der Testbeschreibung entwickelt. Dies erfordert eine Fokussierung auf praktische und fachliche Überlegungen und auch Beschränkungen, die sich aus der Situationsanalyse ergeben und die Testgestaltung und Testdurchführung beeinflussen. Es müssen Entscheidungen über das Format und die Inhalte des Tests gefällt werden; dazu gehört auch, dass der eigene Test mit anderen existierenden Tests zum Sprechen verglichen wird. Wenn der neue Test zum Sprechen als weiterer Subtest in eine schon etablierte Prüfung aufgenommen werden soll, muss sichergestellt werden, dass der neue Prüfungsteil zu den anderen Teilen passt.

Bei der Konzeption des Tests müssen die folgenden Faktoren, die miteinander interagieren, bedacht werden. Es ist wichtig, über diese Interdependenzen Rechenschaft abzulegen und sie zu beschreiben, damit eine Validierung des Tests durchgeführt werden kann. Keine Entscheidung kann unabhängig von anderen getroffen werden; z. B. wird der Umfang und die Komplexität einer Bewertungsanleitung von der Art der gestellten Aufgabe abhängen sowie von den Beurteilern und Prüfern, die zur Verfügung stehen. Wenn viele muttersprachliche Prüfer eingestellt werden können und wenn genug Geld und Zeit zur Verfügung stehen, um sie gründlich zu schulen, ist zu erwarten, dass sie eine umfangreiche und komplexe Bewertungsanleitung anwenden können. Wenn jedoch Zeit und Geld knapp sind und geeignete Prüfer nur schwer gefunden werden können, muss ein Bewertungsverfahren entwickelt werden, das weniger von den Beurteilern fordert.

Während des gesamten Entwicklungsprozesses müssen Entscheidungen als provisorisch angesehen werden und die Bereitschaft vorhanden sein, sich an Änderungen anzupassen und zu früheren Phasen der Entwicklung zurückzukehren. Selbst wenn sich die Prüfung im Echteinsatz befindet, sollte sie nicht als statisch und unveränderbar gelten, da sich im Laufe der Durchführungen noch Bedarf an Veränderung und Verbesserung ergeben wird.

a) Die Kandidaten. Die demographischen Merkmale der Testpopulation müssen ermittelt und ein Kandidatenprofil erstellt werden. Ihr Hintergrund, ihr Alter und ihre Nationalität haben Einfluss auf die Auswahl der Testmaterialien. Das sprachliche Verhalten, das sie zeigen müssen, und die Anwendungsbereiche, in denen ihre Sprachkompetenz getestet werden soll, müssen – zusammen mit dem Niveau ihrer Sprachkompetenz – definiert werden.

b) Die Prüfer. Es muss entschieden werden, welche Qualifikationen die Prüfer für diesen Test vorweisen müssen, welches Training sie erhalten sollen und welche Vorgaben es für ihr Verhalten während des Tests gibt. Möglicherweise werden Sprache, Anweisungen und Fragen der Prüfer im Prüfungsgespräch durch einen festen Rahmen (eine Art Skript für den Prüfer) vorgegeben. Denkbar ist aber auch ein spontaneres Gespräch, besonders wenn es sich um die „Aufwärmphase“ der Prüfung handelt. Die Prüfer müssen in der Bewertungsmethode und der Bewertungsanleitung dieses Tests geschult werden.

c) Die Aufgaben. Sowohl das Format des Tests zum Sprechen insgesamt als auch die Vorgaben, die der Prüfer einsetzt, um bei den Kandidaten eine sprachliche Reaktion hervorzurufen, müssen festgelegt werden. Solche Vorgaben können mündliche oder schriftliche Anweisungen sein oder visuelle Vorgaben wie Fotos, Zeichnungen oder Diagramme. Die Testautoren können möglicherweise aus einer Reihe von Aufgabentypen auswählen, vielleicht müssen sie sich aber auch bei jedem neuen Prüfungsgespräch strikt an die gleiche Folge von Aufgabentypen halten. Es ist auch möglich, dass dem Prüfer mehrere Aufgaben gegeben werden, aus denen er diejenige auswählen kann, die am besten zum jeweiligen Kandidaten passt.

b) Bewertungen. Die sprachlichen Leistungen eines jeden Kandidaten müssen mithilfe einer Bewertungsanleitung entweder während des Tests oder unmittelbar danach bewertet werden. Deshalb muss man eine Balance zwischen der Notwendigkeit komplexer Urteile und der geringen, dafür zur Verfügung stehenden Zeit finden. Vielleicht müssen die Prüfer z. B. jeden Kandidaten auf einer Skala von 1-10 positionieren, bei der jeder mögliche Punktwert durch Deskriptoren definiert wird. Oder die sprachliche Leistung kann hinsichtlich Breite der Strukturen, Korrektheit, Aussprache, Breite des Wortschatzes usw. bewertet werden, wobei jedes dieser Bewertungskriterien nach dem Grad der Beherrschung untergliedert wird, z. B. in die Stufen 1-5. Je komplexer das Bewertungsverfahren ist, desto sorgfältiger müssen die Prüfer ausgesucht und geschult werden.

Fachliche und praktische Überlegungen

In der Konzeptionsphase des Tests müssen sehr viele Aspekte bedacht werden. Man kann sie unterteilen in fachliche und praktische Überlegungen.

a) Fachliche Überlegungen

Fachliche Überlegungen beziehen sich auf das, was genau geprüft werden soll und auf das theoretische Sprachkompetenzmodell, das die Testentwickler ihrer Arbeit zugrunde legen. In der Konzeptionsphase des Tests müssen die realen Situationen bedacht werden, in denen die Kandidaten sich mündlich äußern müssen, sowie das Ausmaß an Kompetenz, das dafür erforderlich ist. Es muss eine Auswahl an möglichen Realsituationen und Sprechanlässen getroffen werden, die der Test abbilden soll. Dabei müssen Kategorien wie Themen, der geforderte Umfang der Lerneräußerung, die konkreten Umstände der Testdurchführung und der Kommunikationskanal (d. h. im direkten Gespräch oder per Telefon) berücksichtigt werden. Auch die Art der Rückmeldung an die Kandidaten über ihre sprachlichen Leistungen muss definiert werden. Man kann ihre Leistungen entweder in Punktwerten oder in Noten ausdrücken oder ihnen nur mitteilen, dass sie den Test bestanden bzw. nicht bestanden haben. Möglich ist auch eine ausführliche Rückmeldung in Form eines Leistungsprofils für jede Aufgabe.

b) Praktische Überlegungen

Alle Aspekte des Tests müssen vom praktischen Standpunkt aus mit den vermutlich zur Verfügung stehenden Mitteln (Zeit, Geld, Personal usw.) geleistet werden können; die vorhandenen Mittel ergeben sich aus der Situationsanalyse. In jeder Testsituation gibt es Beschränkungen und Umstände, an die ein Test angepasst werden muss, die jedoch nicht ideal sind. Die praktischen Überlegungen betreffen die Testdurchführung, die Kandidaten, die Prüfer, die Bewertungen, das Testmaterial, die Aufgaben sowie die Verfahren der Qualitätskontrolle.

I) Die Testdurchführung. Die Überlegungen beinhalten: das vorhandene Personal, um sowohl die Aufgaben zu schreiben als auch bei der Testdurchführung zu helfen; die zur Verfügung stehenden Räume; die für die Durchführung angesetzte Zeit; die Dauer jeder Prüfung und wie schnell die Ergebnisse nach der Prüfung bekannt gegeben werden müssen.

II) Die Kandidaten. Die Überlegungen beinhalten: die Anzahl der Kandidaten, die bewertet werden müssen; wie lange jede Bewertung dauert; ob die Kandidaten einzeln, paarweise oder in größeren Gruppen geprüft werden sollen; ob bei Paar- oder Gruppenprüfungen die Kandidaten ihre(n) Gesprächspartner frei wählen können oder ob die Wahl für sie getroffen wird.

III) Die Prüfer und die Bewertungen. Die Überlegungen beinhalten: die Anzahl der Prüfer, die zur Verfügung stehen; ob die Prüfer alle Muttersprachler sein sollen; wie viele Kandidaten ein Prüfer pro Stunde bewerten soll; ob Prüfer alleine oder zu zweit bewerten sollen (wobei der eine der Prüfer ist und der andere die Bewertung durchführt); wie die Prüfer geschult werden sollen und wie viel Zeit dafür zur Verfügung steht; ob die Vorgaben im Test auf Tonband gegeben werden, falls nicht genügend Prüfer vorhanden sind; und ob dafür ein Sprachlabor zur Verfügung steht.

IV) Die Aufgaben und die Testmaterialien. Die Überlegungen beinhalten: die Anzahl der Phasen oder Testteile, in denen verschiedene Aspekte der Sprechfertigkeit geprüft werden, wobei alle Phasen Eingang in die Bewertung finden müssen; die Aufgabentypen, die in jeder Phase verwendet werden; ob das, was der Prüfer sagt, vorgegeben werden soll oder nicht; welche Art von Vorgaben verwendet werden und wie man verschiedene Aufgabensets zusammenstellt, die identische Anforderungen an die Kandidaten stellen.

V) Die Bewertung. Die Überlegungen beinhalten: wie der erreichte Punktwert den Kandidaten bekannt gegeben wird; wie viele Bewertungen jeder Kandidat erhält; ob die Bewertung holistisch oder analytisch durchgeführt wird (und in letzterem Fall: wie viele getrennte Bewertungskriterien es gibt und wie viele Skalen benutzt werden); welche Verfahren angewendet werden können, um sicherzustellen, dass die Bewertung zuverlässig ist; ob Aufnahmen der mündlichen Prüfung gemacht werden können, um eine zweite oder dritte Bewertung zu ermöglichen, und wer die Bewertungen der Aufnahmen durchführt.

VI) Die Verfahren der Qualitätskontrolle. Die Überlegungen beinhalten: ob Aufnahmen gemacht werden können, um die Qualität der durchgeführten Prüfungen zu kontrollieren; wer diese Kontrollen durchführt; welche anderen Verfahren der Qualitätskontrolle angewendet werden können; wie die Daten gesammelt und für die Analysen und die Validierung archiviert werden und wer diese Analysen und die Validierung durchführen soll.

3. DIE ENTWICKLUNGSPHASE

Am Ende der Konzeptionsphase liegen Beispielmaterialien für den Test vor. Beim Erstellen der Materialien können sich Schwachstellen in der ersten Testbeschreibung zeigen, die daraufhin überarbeitet werden müssen. Wenn dies abgeschlossen ist, beginnt die Entwicklungsphase.

Zu diesem Zeitpunkt werden die Beispielmaterialien sowie Prototypen der Bewertungsanleitung erprobt. Das heißt, dass Lerner mit dem sprachlichen Niveau der Zielgruppe des Tests den Test unter Prüfungsbedingungen ablegen. Die Prüfer, die bei der Erprobung mitarbeiten, können an der Testentwicklung beteiligte Personen sein oder Personen, die sich zum Prüfer für diese Prüfung ausbilden lassen möchten.

Aus Erprobungen kann man eine Menge an Informationen über den Test gewinnen, da die Teilnehmer und die Prüfer detaillierte Rückmeldungen aus ihrer jeweiligen Sicht über viele Aspekte des Tests geben können. Jeder kann seine Reaktionen auf die Textvorlagen, die Themen, die Itemtypen, den Schwierigkeitsgrad, die Länge des Tests usw. schildern. Teilnehmer können beurteilen, wie angemessen die Anweisungen der Prüfer waren und ihre Meinung zu der Einzel- oder Gruppenprüfung sagen, zur Anzahl der eingesetzten Prüfer, und dazu, wie sie die Testsituation insgesamt empfunden haben. Die Prüfer können eine Rückmeldung zum Bewertungssystem geben, das sie angewendet haben. Alle diese Rückmeldungen können in einem Bericht oder mithilfe von Fragebögen gesammelt werden. Testentwickler erhalten weitere Informationen dadurch, dass sie die Durchführung des Tests beobachten, entweder indem sie bei Prüfungsgesprächen dabei sind oder indem sie Prüfungsgespräche aufnehmen (Video- oder Tonaufnahmen). Und schließlich können die Testergebnisse aus der Erprobungsphase zeigen, ob das Testniveau dem Niveau der zukünftigen Kandidaten angemessen ist und die Aufgaben dazu geeignet sind, gute Leistungen hervorzubringen.

Nachdem die Fragebögen, die Skripte der Aufnahmen und die anderen gesammelten Informationen analysiert wurden, werden die Ergebnisse ausgewertet. Der bis dahin stattgefundene Entwicklungsprozess des Tests wird überprüft und sofern erforderlich, werden Änderungen an der Testbeschreibung, den Aufgaben und den Bewertungsanleitungen vorgenommen. Es ist möglich, dass neue Beispielmaterialien geschrieben werden und die Erprobungen und Analysen wiederholt werden müssen; all dies muss im vorgeschriebenen zeitlichen Rahmen erledigt werden. Es kann sein, dass die Erprobungen, Analysen und Auswertungen mehrmals wiederholt werden müssen, bis der Test für den Echteinsatz als geeignet angesehen wird und die Durchführungsphase beginnt.

4. DIE DURCHFÜHRUNGSPHASE

Bevor die Prüfung eingesetzt werden kann, muss man genügend Prüfer und Beurteiler eingestellt und geschult haben, um mit der erwarteten Anzahl an Kandidaten umgehen zu können. Wenn der Test mehr als nur sehr begrenzt eingesetzt wird (z. B. an einer Schule), wird es wahrscheinlich notwendig sein, eine Gruppe von Testautoren zu schulen, die weitere Aufgaben entwickeln, sodass man eine Aufgabenbank aufbauen und bei Bedarf weitere Versionen des Tests zusammenstellen kann. Während der Test wiederholt durchgeführt wird, findet eine Überprüfung statt, um sicherzustellen, dass eine gleichbleibende Schwierigkeit gewahrt wird. Die Daten, die während dieser Überprüfung gesammelt werden, können zu Forschungszwecken verwendet werden. Es sollten Verfahren der Validierung angewendet werden, um nachzuweisen, dass der Test die zu prüfenden Fertigkeiten zuverlässig misst.

Nachdem der Test über einen gewissen Zeitraum eingesetzt wurde, können sich Veränderungen in den Anforderungen ergeben. Die Zielgruppe hat sich evtl. verändert, die Aufgaben sind unter Umständen nicht mehr zeitgemäß und stimmen mit der gängigen Auffassung, wie Tests zum Sprechen aussehen sollten, nicht mehr überein. In diesem Fall wird eine Revision notwendig, die zum Anfang des Zyklus zurückführt und die Entwicklung eines neuen Tests erforderlich macht. Oder man muss zur Konzeptionsphase zurückgehen, d. h., man muss die Testbeschreibung überarbeiten, den Entwicklungsprozess wiederholen und die neuen Materialien erproben.