Universität Duisburg-Essen
Fakultät für Bildungswissenschaften
Lehrstuhl für Lehr-Lernpsychologie

Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz

Dissertation zur Erlangung des Grades Dr. phil.
vorgelegt von
Stefan Hartmann
Dezember 2012

Tag der Disputation:
22. Mai 2013

Gutachter:
Prof. Dr. Detlev Leutner
Prof. Dr. Jürgen Mayer

.

Zusammenfassung

Ziel der vorliegenden Arbeit war die Klärung der Frage, welche Rolle Leseleistungen beim Zustandekommen der Ergebnisse in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenzen spielen. Basierend auf Theorien und Befunden zum multimedialen Lernen (Mayer, 2005b) und zum Verstehen von Texten und Bildern (Schnotz, 2005) wurden drei zentrale Hypothesen aufgestellt, die in zwei Studien mit unterschiedlichen methodischen Ansätzen empirisch geprüft wurden. Zwei der drei Hypothesen konnten bestätigt werden. In einer Regressionsanalyse mit 162 Items aus dem Instrument zur Evaluation der nationalen Bildungsstandards im Fach Biologie (Sek. I) konnte zunächst gezeigt werden, dass der Einsatz von lösungsrelevanten Abbildungen in den Aufgabenstimuli einen signifikanten Anstieg der Lösungswahrscheinlichkeit zur Folge hat. In einer anschließenden experimentellen Studie mit Schülerinnen und Schülern der 9. Klasse ($N = 125$) bestätigte sich dieser Effekt: Bei gleichem Inhalt sind kombinierte Text-Bild-Aufgaben leichter zu lösen als reine Text-Aufgaben. Außerdem konnte nachgewiesen werden, dass sich eine Konfundierung der Ergebnisse mit der Lesegeschwindigkeit der Versuchspersonen durch den Einsatz lösungsrelevanter Abbildungen signifikant reduzieren lässt. Für die Konfundierung der Testleistungen mit Leseverständnis wurde ein schwacher Effekt in der prognostizierten Richtung gefunden, der in der getesteten Stichprobe allerdings nicht signifikant ausfiel. Die Befunde der beiden Studien werden als Beleg dafür gewertet, dass Abbildungen generell geeignet sind, einen Beitrag zur Validität und Fairness von Papier-und-Bleistift-Tests zu leisten.

Papier-und-Bleistift-Tests, Validität, Leseleistungen, Testfairness

Abstract

This thesis examines the role of reading abilities in text-based measures of scientific inquiry skills. Based upon Mayer's Theory of Multimedia Learning (Mayer, 2005b) and Schnotz' Integrated Model of Text and Picture Comprehension (Schnotz, 2005), two studies with different empirical approaches were conducted to investigate the impact of static images on the confounding of reading speed and reading comprehension with the results of a multiple-choice science test. The findings of a regression study with 162 items reveal the positive effect of static images on item difficulty: If images are used to provide relevant information in text-based items, items are significantly less difficult compared to text-only items. In an experimental study ($N = 125$), these findings could be replicated and extended: In addition to the impact on item difficulty, the results show that the use of static images significantly reduces the confounding of reading speed with the measures of the science test. For the reading comprehension, the data indicates a weak interaction as well, but no significant effect was found in the tested sample. However, the findings of both studies suggest that static images can be used to increase a test's validity and fairness.

paper-and-pencil tests, validity, reading abilities, test fairness

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Ausgangssituation und Forschungsfragen

Wenn im Rahmen von Schulleistungsstudien naturwissenschaftliche Kompetenzen von Schülerinnen und Schülern erfasst werden, geschieht dies in der Regel vorrangig mit Papier-und-Bleistift-Tests. Im Vergleich zu anderen Verfahren wie Hands-On-Aufgaben oder computerbasierten Assessments stellen schriftliche Instrumente eher geringe Ansprüche an Personal und Infrastruktur und ermöglichen eine reliable Testung bei vergleichsweise geringen Kosten (Haldane, 2009; Stecher & Klein, 1997). Die Versuchspersonen haben meist in Textform dargebotene Aufgaben zu bearbeiten, in denen offene, halboffene oder geschlossene Antwortformate zum Einsatz kommen. Bei vertiefenden Analysen der Ergebnisse lassen sich mittlere bis hohe Korrelationen zwischen den erhobenen Leistungen in Naturwissenschaften und Lesekompetenz finden (Leutner, Klieme, Meyer & Wirth, 2004; Organisation for Economic Cooperation and Development [OECD], 2003, 2009). Dies ist kaum verwunderlich, denn logischerweise wird zum Bearbeiten textbasierter Aufgaben immer ein gewisses Maß an Lesefähigkeiten erforderlich sein.

Bei der Diskussion möglicher Zusammenhänge zwischen sprachlichen Fähigkeiten und den Testleistungen in Large-Scale-Assessments wird zum Teil auf die Rolle der Sprache beim Kompetenzerwerb verwiesen (Bonsen, Kummer & Bos, 2008; Paetsch & Radmann, 2011), zum Teil aber auch auf die sprachlichen Anforderungen beim Bearbeiten der Testaufgaben selbst (Leutner et al., 2004; Prenzel, Häußler, Rost & Senkbeil, 2002). Während Testentwickler einräumen,

dass in schriftlichen Aufgaben eine gewisse Konfundierung mit Lesefähigkeiten nicht zu vermeiden ist (z. B. Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010; Martinello, 2009), nehmen einige Kritiker die Befunde gar zum Anlass, die Konstruktvalidität der betreffenden Instrumente generell infrage zu stellen (Klein, 2010; Rindermann, 2006).

Die Frage, ob die oben genannten Zusammenhänge zwischen Naturwissenschafts- und Lesetests tatsächlich auf die sprachlichen Anforderungen beim Bearbeiten der Aufgaben zurückgehen, lässt sich auf der Basis bisheriger Studien nicht sicher beantworten. An dieser "Informationslücke" setzt die vorliegende Dissertation an. Aus psychometrischer Sicht stellen sich dabei folgende allgemeine Forschungsfragen:

1. Welche Rolle spielen individuelle Leseleistungen beim Zustandekommen der Ergebnisse in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenzen?

2. Müssen Leseleistungen im Kontext schriftlicher Instrumente zur Erfassung naturwissenschaftlicher Kompetenzen als konfundierende Variablen berücksichtigt werden—und falls ja: Wie stark sind die Konfundierungseffekte?

3. Welche Itemmerkmale sind ggf. für die Stärke der Konfundierungseffekte bedeutsam?

4. Welche Personenmerkmale sind ggf. für die Stärke der Konfundierungseffekte bedeutsam?

1.2 Ziele der Studie

Um Antworten auf die vorgenannten Fragen zu finden, soll in der vorliegenden Dissertation die Rolle von Leseleistungen beim Zustandekommen der Ergebnisse in schriftlichen Tests zur Erfassung von naturwissenschaftlichen Kompetenzen mithilfe quantitativer Methoden genauer untersucht werden. Basierend auf Theorien und Befunden zum multimedialen Lernen (Mayer, 1997, 2005a) und zum Verstehen von Texten und Bildern (Kürschner & Schnotz, 2008; Perfetti, 1985; Schnotz, 2005, 2006; van Dijk & Kintsch, 1983) sollen insbesondere die Effekte unterschiedlicher Aufgabenformate—hier im Sinne der verwendeten Darbietungsmodalität in den Testaufgaben—auf die Ergebnisse im Bereich

Naturwissenschaften und auf die Zusammenhänge dieser Ergebnisse mit Leseleistungen analysiert werden. Dabei gilt es Wege zu finden, den zum Lösen erforderlichen Leseaufwand zu reduzieren, ohne dass sich dabei Einschränkungen im Informationsgehalt der Aufgaben ergeben. Neben gedrucktem Text werden statische Abbildungen als alternatives Format zur Darbietung lösungsrelevanter Informationen in den Aufgabenstimuli zum Einsatz kommen. Über die Lesefähigkeiten hinaus sollen weitere Moderatorvariablen, die für das Zustandekommen der Ergebnisse möglicherweise bedeutsam sind, im Rahmen einer Vorstudie identifiziert und in die Hauptuntersuchung einbezogen werden.

1.3 Kontext

Die Erhebung der zur Beantwortung der Forschungsfragen notwendigen Daten erfolgt in enger Anlehnung an die Evaluation der Bildungsstandards in den Naturwissenschaften in der Sekundarstufe I (ESNaS) in Deutschland (Walpuski et al., 2010). Um ein reliables und ökonomisches Instrument zur Überprüfung der Hypothesen dieser Arbeit zu erhalten, werden Testitems aus dem vorhandenen ESNaS-Aufgabenpool entnommen und teilweise umkonstruiert. Dabei ist eine Beschränkung auf bestimmte Themenbereiche bzw. Subskalen des Kompetenzmodells von ESNaS geplant. Zur Operationalisierung naturwissenschaftlicher Kompetenzen werden Aufgaben zur Anwendung kommen, die zur Erfassung von Fähigkeiten der naturwissenschaftlichen Erkenntnisgewinnung im Fach Biologie entwickelt wurden.

Zum Zeitpunkt der Erstellung der vorliegenden Arbeit sind so gut wie keine Ergebnisse aus dem ESNaS-Projekt publiziert. Um bei der Formulierung zentraler Hypothesen dennoch nicht auf die notwendige empirische Basis verzichten zu müssen, wird zum einen auf Befunde vergleichbarer Large-Scale-Assessments zurückgegriffen, insbesondere die Ergebnisse aus den PISA-Erhebungen sowie deren nationaler Ergänzungsstudien. Zum anderen werden eigene Berechnungen auf Basis von Daten aus der Pilotierungsstudie des ESNaS-Projektes angefertigt.

1.4 Wissenschaftliche und gesellschaftliche Relevanz

Die Ergebnisse sollen einen allgemeinen Beitrag zur Diskussion über das Zustandekommen der Leistungen in schriftlichen Tests und über deren Validität und Testfairness leisten. Diese Diskussion ist nicht nur von pädagogisch-psychologischer, sondern auch von bildungspolitischer und gesellschaftlicher Relevanz (Bundesministerium für Bildung und Forschung [BMBF], 2007a; Weinert, 2001). So hängt die Akzeptanz, die großflächigen Schulleistungsstudien und den auf Basis der Ergebnisse getroffenen Entscheidungen entgegengebracht wird, nicht zuletzt von der Güte ab, mit der diese Instrumente die anvisierten Fähigkeiten messen. Eine entsprechende Verallgemeinerbarkeit der Befunde dieser Dissertation ist insofern gegeben, als die zugrundeliegenden Kompetenzbegriffe und die Art der Instrumente in allen Large-Scale-Assessments gewisse Gemeinsamkeiten aufweisen.

Daneben sollen die Ergebnisse aber auch speziell zur Validierung des Instrumentes zur Evaluation der Standards in den Naturwissenschaften in der Sekundarstufe I beitragen. Sie sollen Antworten auf die Frage geben, inwieweit sich bestimmte Aspekte der Gestaltung von Testaufgaben auf die Validität des Tests ausgewirkt haben. Dabei steht vor allem die mögliche Konfundierung des Instruments mit Leseleistungen im Fokus.

Vor dem Hintergrund der Testfairness (Bortz & Döring, 2006, S. 192 f.) soll die systematische Untersuchung von Interaktionen aus Darbietungsformat und bestimmten Personenvariablen zudem helfen, Möglichkeiten für eine gerechtere Testung bestimmter Personengruppen auszuloten.

1.5 Struktur der Arbeit

Den Schwerpunkt dieser Dissertation bildet eine Experimentalstudie, in deren Rahmen die lösungsrelevanten Informationen ausgewählter Aufgaben aus der Evaluation der Standards in den Naturwissenschaften in zwei unterschiedliche Darbietungsformate (Text und statische Abbildungen) überführt wurden. Basierend auf diesen beiden Formaten wurden zwei parallele Testformen entwickelt, die im Sekundarschulbereich eingesetzt und anschließend mithil-

fe quantitativer Methoden verglichen wurden. Die Umkonstruktion erforderte zwei Vorstudien, um ein Instrument zu erhalten, das in beiden Parallelformen eine zufriedenstellende interne Konsistenz der Skalen aufwies. Da den Vorstudien und der Hauptuntersuchung dieselben theoretischen Annahmen zugrunde liegen, bildet eine allgemeine theoretische Einführung den Beginn der Arbeit. Darauf folgt eine Analyse von Daten aus dem ESNaS-Projekt. Anschließend werden die Vorstudien und die Hauptstudie in separaten Abschnitten vorgestellt, die jeweils über einen eigenen Methoden- und Ergebnisteil sowie eine Diskussion der Befunde verfügen. Zum Ende der Arbeit werden die Resultate aller Studien zusammengefasst und ihre theoretischen und praktischen Implikationen diskutiert.

Kapitel 2

Theoretischer Hintergrund und zu prüfende psychologische Hypothesen

Zu Beginn dieser Arbeit ist zunächst zu klären, warum überhaupt schriftliche Aufgaben zur Erfassung naturwissenschaftlicher Kompetenzen eingesetzt werden und somit eine Konfundierung mit Leseleistungen "riskiert" wird. Dazu soll kurz auf die allgemeine Konzeption von Educational Large-Scale-Assessments eingegangen werden. Anschließend werden naturwissenschaftliche Kompetenz, verschiedene Konstrukte zur Beschreibung und Erfassung von Leseleistungen und die Zusammenhänge zwischen diesen Variablen anhand bisheriger Theorien und Befunde definiert und diskutiert sowie die im Rahmen dieser Arbeit zu prüfenden Hypothesen aufgestellt.

2.1 Kompetenzmessung im Large-Scale-Assessment

Die leistungsindizierte Evaluation von Bildungssystemen findet in der Regel in Form großflächig angelegter, quantitativer Schulleistungsstudien, sogenannter Educational Large-Scale-Assessments, statt. Deren Ergebnisse haben sich sowohl in Deutschland als auch international zu einem wesentlichen Element einer indikatorenorientierten Bildungsberichterstattung entwickelt und ermög-

lichen unter anderem eine präzise Einschätzung von Kompetenzniveaus und Kompetenzunterschieden auf föderaler, nationaler oder internationaler Ebene, Vergleiche unterschiedlicher Schultypen oder die Beurteilung von Zusammenhängen zwischen sozioökonomischen bzw. soziokulturellen Variablen und Bildungschancen. Die Ergebnisse von Large-Scale-Assessments sind nicht nur als Forschungsgegenstand in den Bildungswissenschaften und als Informationsquelle für die Unterrichtsgestaltung und das Schulwesen, sondern auch als Entscheidungsgrundlage für die Politik interessant (BMBF, 2007b; Prenzel & Baumert, 2008; Weinert, 2001).

Zu den in Wissenschaft, Bildungspolitik, Medien und Gesellschaft am meisten beachteten und diskutierten Large-Scale-Assessments zählen vor allem die von der Organisation for Economic Cooperation and Development (OECD) initiierten PISA-Studien und die von der International Association for the Evaluation of Educational Achievement (IEA) durchgeführten Studien TIMSS und PIRLS / IGLU. Daneben gibt es auch eine Reihe nationaler Erhebungen wie das seit 1969 in den USA durchgeführte National Assessment of Educational Progress (NAEP). In der Bundesrepublik haben insbesondere die Ergebnisse der internationalen Studien PISA, TIMSS und PIRLS (sowie deren nationale Ergänzungsstudien) eine hohe Medienwirkung entfaltet und zu regen fachdidaktischen und bildungspolitischen Debatten geführt (Weinert, 2001). Im Rahmen der politischen Auseinandersetzung mit der Thematik hat sich die Ständige Konferenz der Kultusminister der Länder (KMK) auf bundesweit gültige Bildungsstandards für eine Reihe von Unterrichtsfächern geeinigt und deren regelmäßige empirische Überprüfung beschlossen (BMBF, 2007a). Mit dem Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin wurde eine wissenschaftliche Einrichtung gegründet, deren zentrale Aufgabe die Evaluation dieser Bildungsstandards ist (Ständige Konferenz der Kultusminister der Länder / Institut zur Qualitätsentwicklung im Bildungswesen [KMK / IQB], 2006).

2.1.1 Kompetenzen

Obwohl sich die vorgenannten Studien in vielen Punkten unterscheiden, gibt es eine Reihe von Gemeinsamkeiten. Primäres Ziel von PISA, TIMSS, PIRLS, NAEP und der Evaluation der Bildungsstandards ist zunächst—ganz allge-

mein—die Messung bestimmter Fähigkeiten und Fertigkeiten von Schülerinnen und Schülern. Dabei werden in der Regel kognitive Leistungen erfasst, die betont funktional ausgerichtet sind und unter dem Begriff *Kompetenzen* zusammengefasst werden. In Anlehnung an das im angelsächsischen Raum verbreitete Konstrukt der *Literacy* weisen Kompetenzen eine deutliche Nähe zur Bewältigung von Problemen in authentischen Anwendungssituationen auf. Gemäß einer häufig zitierten Definition sind Kompetenzen "die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten oder Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (Weinert, 2001, S. 27 f.). Andere Definitionen beschränken den Kompetenzbegriff auf kognitive Leistungsdispositionen und schließen sogenannte Handlungskompetenzen, die motivationale, volitionale und soziale Aspekte einschließen, aus (Hartig & Klieme, 2006; Klieme & Leutner, 2006).

Im Kontext der vorgenannten Large-Scale-Assessments werden vorrangig Kompetenzen getestet, deren Erwerb zu den zentralen Bestandteilen schulischer Bildung gezählt wird, beispielsweise Kompetenzen in Mathematik, im Lesen oder in den Naturwissenschaften. Die Orientierung an nationalen oder länderspezifischen Curricula fällt in den einzelnen Studien unterschiedlich stark aus.

2.1.2 Instrumente

Die Erfassung von Kompetenzen in Large-Scale-Assessments ist hinsichtlich ihrer Methodik von einigen Besonderheiten gekennzeichnet. Wenngleich sich die Instrumente und Methoden von PISA, TIMSS und der Evaluation der Bildungsstandards in vielerlei Hinsicht unterscheiden, so ist ihnen doch gemein, dass sie vorrangig das Bearbeiten textbasierter Aufgaben als zentrale Methode der Leistungsindikation nutzen.

Um zuverlässige und belastbare Aussagen (beispielsweise über Unterschiede der Bildungssysteme von Bundesländern oder Staaten) treffen zu können, ist zudem die Ziehung sehr großer Stichproben notwendig. An internationalen Large-Scale-Assessments nehmen in der Regel mehrere zehn- oder hunderttausend Schülerinnen und Schüler teil. Als Beispiel sei hier die PISA-Erhebung

von 2006 genannt, in der die Leistungen von knapp 400 000 Personen evaluiert wurden (OECD, 2009, S. 22). Um eine derart große Anzahl von Schülerinnen und Schülern in einem zeitlich und materiell überschaubaren Rahmen testen zu können, bedarf es möglichst ökonomischer Instrumente. Meistens kommen dabei Papier-und-Bleistift-Tests zum Einsatz, da diese eine reliable Testung bei vergleichsweise geringen Kosten ermöglichen (Stecher & Klein, 1997).

Andere Formen der Erfassung von Kompetenzen sind computerbasierte Assessments und praktische Tests. Erstere können zwar einen Beitrag zur Verbesserung der Testökonomie leisten (Jude & Wirth, 2007), setzen dabei allerdings eine gewisse Infrastruktur voraus, die bei großen internationalen Erhebungen in der Regel nicht gewährleistet werden kann und die im Falle einer Neuanschaffung erhebliche Kosten verursacht (Haldane, 2009). Zur Messung von praktischen Fähigkeiten wie dem Durchführen von Versuchen oder dem Umgang mit Hilfsmitteln (z. B. einem Mikroskop) eignen sich praktische Tests (sog. *Hands-On Tasks*). Sie weisen im Vergleich zu schriftlichen Aufgaben unter Umständen eine größere Realitätsnähe auf. Es ist jedoch nicht zweifelsfrei geklärt, ob sie dadurch tatsächlich auch die valideren Instrumente darstellen (Baker, O'Neil & Linn, 1993). Hinsichtlich der Reliabilität reichen praktische Tests nicht an die Güte heran, die mit klassischen Papier-und-Bleistift-Aufgaben erreicht wird (Stecher & Klein, 1997). Zudem gehen sie mit einem deutlich höheren personellen und zeitlichen Aufwand einher und sind nicht zuletzt deshalb um ein Vielfaches teurer als schriftliche Instrumente (ebd.).

Trotz eines zunehmenden Einsatzes computerbasierter Instrumente zur Kompetenzdiagnostik (Jude & Wirth, 2007; Scheuermann & Björnsson, 2009) werden bei der Kompetenzmessung in Educational Large-Scale-Assessments immer noch vorwiegend Papier-und-Bleistift-Tests verwendet. Diese Tests bestehen in der Regel aus mehreren schriftlich dargebotenen Aufgabenstimuli, an die sich jeweils ein oder mehrere Testitems anschließen. Aufgabenstimuli und Testitems enthalten neben reinem Text häufig auch Informationen in anderer Form, z. B. Abbildungen, Tabellen oder Diagramme.

Die Lösung der Items erfolgt in Form von Markierungen, Zeichnungen oder schriftlichen Antworten im Testheft oder auf separaten Antwortbögen. Hinsichtlich der Antwortformate werden dabei offene (auch "freie") und geschlossene (auch "gebundene") Antworten unterschieden. Die Erfassung von Fähig-

keiten über offene Antworten bietet sich beispielsweise an, "wenn die Lösungsfindung selbsttätig erfolgen soll, ... der Lösungsweg interessiert [oder] das Erraten der richtigen Lösung oder das Wählen der nächstbesten Antwort verhindert werden soll" (Lienert & Raatz, 1998, S. 25). Demgegenüber bieten geschlossene Antworten im Vergleich zu offenen Formaten u. a. Vorteile aufgrund ihrer kürzeren Bearbeitungszeit und ihrer einfacheren und objektiveren Auswertbarkeit (Lienert & Raatz, S. 26).

Während das Beantworten offener Antwortformate in der Regel eine aktive sprachliche Leistung (und zwar das selbständige Formulieren von Wörtern, Wortgruppen oder Sätzen) darstellt und somit höchstwahrscheinlich mit der schriftlichen Artikulationsfähigkeit und -bereitschaft der Versuchspersonen konfundiert ist (Lienert & Raatz, 1998, S. 26), erfordern geschlossene Aufgaben hinsichtlich der sprachlichen Fähigkeiten nur das "passive" Lesen und Verstehen des Aufgabenstimulus' und der Antwortalternativen, da letztere vorgegeben sind und durch Ankreuzen markiert werden. Hinsichtlich der Ökonomie und Reliabilität sind Ankreuz-Formate offenen Antworten deutlich überlegen (Stecher & Klein, 1997, S. 10). Über die Frage, ob das Beantworten offener und geschlossener Items unterschiedliche Fähigkeiten erfordert, besteht kein Konsens (vgl. Hollingworth, Beard & Proctor, 2007).

In vielen Tests kommt—nicht zuletzt aus Gründen der Testökonomie—ein Mix der Antwortformate zum Einsatz (vgl. OECD, 2009, S. 22; Olson, Martin & Mullis, 2008, S. 23). Bei der Konstruktion von Items für die Evaluation der Bildungsstandards in den Naturwissenschaften wird in etwa eine Verteilung von 50 % Ankreuz-Antworten, 30 % halboffenen Antworten (z. B. Lückentexte oder Formelergänzungen) und 20 % offenen Antworten anvisiert.

2.2 Schriftliche Erfassung naturwissenschaftlicher Kompetenzen

Das Deutsche PISA–Konsortium (2004) definiert die naturwissenschaftliche Grundbildung als "die Fähigkeit, naturwissenschaftliches Wissen anzuwenden, naturwissenschaftliche Fragen zu erkennen und aus Belegen Schlussfolgerungen zu ziehen, um Entscheidungen zu verstehen und zu treffen, die die natürliche Welt und die durch menschliches Handeln an ihr vorgenommenen Ver-

änderungen betreffen" (S. 112). Naturwissenschaftliche Grundbildung ist eine Voraussetzung für die "aktive Teilhabe an gesellschaftlicher Kommunikation und Meinungsbildung über technische Entwicklung und naturwissenschaftliche Forschung und ist deshalb wesentlicher Bestandteil von Allgemeinbildung" (KMK, 2005a, S. 6). Ein Verständnis naturwissenschaftlicher Fragen, Inhalte und Methoden sowie die kritische Reflexion über deren Bedeutung für das eigene Handeln sind in unserer Gesellschaft unverzichtbar (Wittwer, Saß & Prenzel, 2008, S. 87). Aus diesen Gründen sind naturwissenschaftliche Bildung, naturwissenschaftlicher Unterricht und naturwissenschaftliche Kompetenzen das Thema zahlreicher Untersuchungen und Veröffentlichungen[1] sowie fester Bestandteil nationaler und internationaler Large-Scale-Assessments (Bos et al., 2008; PISA-Konsortium Deutschland, 2007).

Um zu gewährleisten, dass naturwissenschaftliche Kompetenzen in der schulischen Bildung umfassend vermittelt werden, hat sich die Kultusministerkonferenz mit ihrem Beschluss vom 16.12.2004 auf einheitliche Bildungsstandards für den mittleren Schulabschluss in den Fächern Biologie, Chemie und Physik geeinigt (KMK, 2005a, b, c). Diese Standards sind seit 2005 verbindlich und sollen regelmäßig überprüft werden (Kauertz et al., 2010). Da die vorliegende Arbeit im Rahmen eines interdisziplinären Forschungsvorhabens zur Evaluation dieser Bildungsstandards angefertigt wird, soll das Projekt im Folgenden kurz umrissen werden.

2.2.1 Evaluation der Bildungsstandards in den Naturwissenschaften

In den Bildungsstandards für die Fächer Biologie, Chemie und Physik sind Fähigkeiten und Fertigkeiten beschrieben, über welche Schülerinnen und Schüler am Ende der Sekundarstufe I verfügen sollen. Diese "gehen auf die Definition von Kompetenz nach Weinert (2001) zurück" (Kauertz et al., 2010, S. 136), beschränken sich dabei jedoch überwiegend auf kognitive Leistungsdispositionen, die im schulischen Kontext vermittelt werden, wohingegen motivationale und volitionale Aspekte weitgehend unberücksichtigt bleiben. Die Evaluation der

[1] Eine Datenbanksuche in PsycINFO ergab zum Zeitpunkt der Verschriftlichung dieser Arbeit 8 626 Treffer für den Suchbegriff "science education", 2 249 für "science teaching" und 179 für "scientific literacy" (Abfrage vom 15. August 2011).

Bildungsstandards erfolgt in einer Kooperation von Fachdidaktikerinnen und Fachdidaktikern, Psychometrikern, linguistischen Bewerterinnen und Bewertern sowie erfahrenen Lehrkräften aus dem gesamten Bundesgebiet (Walpuski et al., 2010).

Trotz unterschiedlicher Fachinhalte verbindet die Naturwissenschaften eine Reihe von Gemeinsamkeiten, sodass sich die fachdidaktischen Leitungen für Biologie, Chemie und Physik auf ein einheitliches Kompetenzmodell für alle drei Fächer geeinigt haben. Das Modell unterscheidet innerhalb des jeweiligen Faches die vier Kompetenzbereiche *Fachwissen*[2], *Erkenntnisgewinnung*, *Bewertung* und *Kommunikation* (KMK, 2005a, b, c). Diese Kompetenzbereiche sind jeweils in diverse Teilbereiche und Aspekte untergliedert.

Die in der hier vorliegenden Arbeit eingesetzten Aufgaben entstammen dem Kompetenzbereich *Erkenntnisgewinnung* im Fach Biologie. Zu diesem Themenkomplex liegen bisher deutlich weniger Erkenntnisse vor als zum Fachwissen.[3]

2.2.2 Kompetenzmodell zur naturwissenschaftlichen Erkenntnisgewinnung

Der Bereich *Erkenntnisgewinnung* beschreibt die wissenschaftsmethodischen Kompetenzen von Schülerinnen und Schülern. Diese Kompetenzen gehören zum Kern der naturwissenschaftlichen Bildung und umfassen ein breites Spektrum von Fähigkeiten (Bybee, 2002, S. 33f.). Bei der Evaluation der Bildungsstandards werden drei Kompetenzteilbereiche der Erkenntnisgewinnung unterschieden: *naturwissenschaftliche Untersuchungen*, *naturwissenschaftliche Modellbildung* und *wissenschaftstheoretische Reflexion*.

[2] Um zu verdeutlichen, dass bei der Modellierung des Bereichs *Fachwissen* kein Wissen, sondern Kompetenzen erfasst werden, wurde der Bereich im Rahmen von ESNaS in *Umgang mit Fachwissen* umbenannt.

[3] Eine Datenbanksuche in PsycINFO ergab 1 976 Treffer für "scientific knowledge" und 861 Treffer für "scientific inquiry". In Kombination mit dem Stichwort "assessment" ergaben sich 144 Treffer für "scientific knowledge" und 82 für "scientific inquiry" (Abfrage vom 15. August 2011).

Naturwissenschaftliche Untersuchungen.

Wenngleich die Gewinnung naturwissenschaftlicher Erkenntnisse in hohem Maß an praktische Methoden wie das Beobachten, Vergleichen und Experimentieren gebunden ist und diese Methoden auch im Unterricht eine wichtige Rolle einnehmen, werden die entsprechenden Kompetenzen von Schülerinnen und Schülern im Rahmen des ESNaS-Projekts als kognitive Fähigkeiten modelliert und evaluiert. Der hypothetisch-deduktive Erkenntnisprozess ist dabei als eine domänenspezifische Form des Problemlösens definiert (Helgeson, 1993; Mayer, 2007). Die zentralen Schritte dieses Prozesses konnten als vier eigenständige Dimensionen identifiziert und empirisch abgebildet werden (Grube, 2010; Grube, Möller & Mayer, 2007). Diese Dimensionen wurden für die Beschreibung des Kompetenzbereichs *Erkenntnisgewinnung* in das Projekt ESNaS übernommen. Es handelt sich dabei um:

- Das Formulieren einer naturwissenschaftlichen *Fragestellung*.
- Das Aufstellen einer naturwissenschaftlichen *Hypothese*.
- Die *Planung* einer Untersuchung zur Prüfung der Hypothese.
- Die *Auswertung und Interpretation* der erhobenen Daten.

Die praktische Durchführung einer geplanten Untersuchung bleibt in dieser Auflistung scheinbar unberücksichtigt. Praktischen Arbeitstechniken wird aber sowohl im naturwissenschaftlichen Unterricht als auch in den Kompetenzbeschreibungen der Bildungsstandards eine große Bedeutung beigemessen. Dies wird an diversen Standards aus dem Bereich *Erkenntnisgewinnung* deutlich (KMK, 2005a, b, c). Beispielhaft sei hierfür der Standard E1 aus dem Fach Biologie genannt: “Die Schülerinnen und Schüler ... mikroskopieren Zellen und stellen sie in einer Zeichnung dar” (KMK, 2005a, S. 14). Zur Messung derartiger Fähigkeiten wären streng genommen praktische Tests, sogenannte Hands-On-Aufgaben, erforderlich. Im Rahmen von Papier-und-Bleistift-Tests lassen sich solche Kompetenzen hingegen nicht unmittelbar prüfen. Um dieser Tatsache bei der Evaluation der Bildungsstandards hinreichend Rechnung zu tragen, mussten entsprechende Kompetenzformulierungen aus den Standards in einem weiteren Sinne interpretiert werden (Kauertz et al., 2010).

Naturwissenschaftliche Modellbildung.

Neben dem Planen, Durchführen und Auswerten von Untersuchungen ist das naturwissenschaftliche Arbeiten durch den Umgang mit Modellen gekennzeichnet. Modelle lassen sich "einerseits als Medien auffassen, die die Vermittlung von naturwissenschaftlichen Kenntnissen unterstützen sollen, oder andererseits als Denk- und Arbeitsmethoden nutzen, um im Prozess der Erkenntnisgewinnung eingesetzt zu werden" (Upmeier zu Belzen & Krüger, 2010, S. 42). Unter Modellen werden somit sowohl gegenständliche Struktur- oder Funktionsmodelle als auch abstrakte oder mathematische Modelle und Theorien verstanden. Als Kompetenzteilbereich im Projekt ESNaS umfasst die Modellbildung Kenntnisse der Funktionalität von Modellen, die Anwendung geeigneter Modelle beim wissenschaftlichen Arbeiten sowie die Beurteilung der Grenzen von Modellen. Der Prozess der Modellbildung besitzt Parallelen zum naturwissenschaftlichen Untersuchen (Mayer, 2007), was sich bei der Evaluation der beiden Bereiche in gewissen Ähnlichkeiten der Aufgabenstellungen niederschlägt.

Wissenschaftstheoretische Reflexion.

Die Gewinnung naturwissenschaftlicher Erkenntnisse sollte logisch-deduktiven Prinzipien folgen (Popper, 1994; Popper & Miller, 1983). Der Erkenntnisprozess ist idealerweise durch Falsifizierbarkeit, Intersubjektivität, Reproduzierbarkeit, Wertfreiheit und Widerspruchsfreiheit gekennzeichnet (Mayer, Harms, Hammann, Bayrhuber & Kattmann, 2004). Naturwissenschaftliches Wissen hat immer einen vorläufigen Charakter und ist häufig auch durch gesellschaftliche oder technologische Aspekte geprägt. Die Kenntnis dieser Eigenschaften ist in ESNaS unter einem dritten Kompetenzteilbereich zusammengefasst, der als *Wissenschaftstheoretische Reflexion* bezeichnet wird (Walpuski et al., 2010).

Die Aufgaben zur Evaluation dieses Bereichs sind eher auf einer Metaebene des Wissenschaftsverständnisses angesiedelt und unterscheiden sich deutlich von den Aufgaben zu den beiden anderen Bereichen. In der vorliegenden Untersuchung werden deshalb nur Aufgaben aus den ersten beiden Bereichen (*naturwissenschaftliche Untersuchungen* und *naturwissenschaftliche Modellbildung*) zum Einsatz kommen, jedoch keine Aufgaben zum Bereich *wissenschaftstheoretische Reflexion*.

2.2.3 Schwierigkeitserzeugende Aufgabenmerkmale

Um bereits ab dem Beginn der Konstruktion von Testaufgaben zur Evaluation der Bildungsstandards sicherzustellen, dass den Fähigkeiten der Schülerinnen und Schüler entsprechend ein ausreichend großes Spektrum an Aufgabenschwierigkeiten abgedeckt wird, wurde im Projekt ESNaS die Verteilung zweier schwierigkeitserzeugender Itemmerkmale systematisch kontrolliert. Es handelt sich dabei um die Komplexität der zur Lösung zu verarbeitenden Informationen sowie um den zum Lösen erforderlichen kognitiven Prozess.

Hinsichtlich der Komplexität werden fünf "Stufen" unterschieden, bei denen jeweils von einem Niveau zum nächsten ein Ansteigen der Aufgabenschwierigkeit erwartet wird (Kauertz et al., 2010):

- *ein Fakt*
- *zwei Fakten*
- *ein Zusammenhang*
- *zwei Zusammenhänge*
- *übergeordnetes Konzept*

Bei den zum Lösen erforderlichen kognitiven Prozessen wird ebenfalls davon ausgegangen, dass sie schwierigkeitserzeugend sind (Kauertz et al., 2010). In aufsteigender Schwierigkeit sind das die kognitiven Prozesse:

- *Reproduzieren*
- *Selegieren*
- *Organisieren*
- *Integrieren*

Das Reproduzieren erfordert die reine Wiedergabe einer vorgegebenen Information. Beim Selegieren müssen die Schülerinnen und Schüler die richtige Information aus mehreren Möglichkeiten auswählen. Der Prozess des Organisierens beinhaltet das richtige Anordnen von Informationen in linearen Abfolgen, Hierarchien oder Kreisläufen. Wenn zum Lösen einer Aufgabe der Transfer von

Informationen auf eine völlig neue Situation erforderlich ist, wird von Integrieren gesprochen. Bei der Konstruktion von Aufgaben für das Projekt ESNaS ist eine ungefähre Verteilung von 35 % (Reproduzieren), 25 % (Selegieren), 25 % (Organisieren) und 15 % (Integrieren) anvisiert. Die tatsächliche Verteilung kann unter Umständen von dieser Vorgabe abweichen (Wellnitz, 2012).

Insbesondere die kognitiven Prozesse *Reproduzieren* und *Selegieren*, die zusammen einen Großteil der konstruierten Aufgaben ausmachen, zeichnen sich in erster Linie durch die reine Entnahme von Informationen aus den schriftlich dargebotenen Aufgabenstimuli aus und weisen somit eine gewisse Nähe zu gängigen Konzepten von Leseverständnis auf (vgl. Kap. 2.3). Dass diese Aufgaben vermutlich weniger hohe Anforderungen an fachliche oder methodische Kompetenzen stellen und stattdessen in größerem Maß grundlegende Fähigkeiten der Informationsentnahme testen, ist unter anderem dem Wunsch geschuldet, Defizite auf den unteren Kompetenzniveaus differenzierter zu erfassen, als dies in bisherigen Studien der Fall war. Gemäß Kauertz et al. (2010) ist "[i]nsbesondere im unteren Leistungsbereich ... ein hoher Zusammenhang zur Lesekompetenz zu erwarten, da diese den Erwerb und die Prüfung naturwissenschaftlicher Kompetenzen erst ermöglicht" (S. 150).

Aber auch in den höheren Schwierigkeitsbereichen bildet die Entnahme schriftlicher Informationen aus dem Stimulus eine notwendige Voraussetzung für das erfolgreiche Lösen der Aufgaben. Es stellt sich deshalb die Frage, welche Rolle Leseleistungen beim Zustandekommen der Ergebnisse generell spielen. Um fundierte Hypothesen zu dieser Frage aufstellen zu können, bedarf es zunächst einer klaren Definition für Leseleistungen.

2.3 Lesen

Schrift stellt eines der ältesten Verfahren zur Kodierung und dauerhaften Aufbewahrung von Informationen dar. Das früheste, als Alphabetschrift bezeichnete Zeichensystem wird der phönizischen Kultur zugeschrieben und ist auf das zweite vorchristliche Jahrtausend datiert (Haarmann, 1991, S. 268). Die zu Beginn unseres Jahrhunderts am weitesten verbreitete Schrift—das lateinische Alphabet—ist heute schätzungsweise 2 700 Jahre alt (Haarmann, 1991, S. 294). Allen Schriften ist gemein, dass sie das Kodieren sprachlicher Informa-

tionen in visuell wahrnehmbare Zeichen ermöglichen. Während sich entwicklungsgeschichtlich ältere Schriften an der Wortbedeutung orientieren (Logographie), ist bei später ausgebildeten Schriften die Differenzierung der Laute für die Anwendung einzelner Schriftzeichen ausschlaggebend (Phonographie) (Haarmann, 1991, S. 148). Die Dekodierung von Schriftzeichen wird als Lesen bezeichnet und erfordert im Fall phonographischer Schriftsysteme sowohl ein Verständnis des verwendeten Zeichensystems (z. B. lateinische, kyrillische oder arabische Schrift) als auch das Beherrschen der verwendeten Sprache (z. B. Deutsch, Russisch oder Hocharabisch).

Lesen ist eine kognitive Fähigkeit und als solche seit vielen Jahren Gegenstand zahlreicher wissenschaftlicher Untersuchungen.[4] In der Leseforschung existieren verschiedene Paradigmen, die von der Lesefähigkeit über das Textverstehen bis zur Lesekompetenz reichen.

2.3.1 Lesefähigkeit

Die Fähigkeit, Grapheme in Phoneme umzuwandeln und somit geschriebene Texte zu dekodieren, wird als Lesefähigkeit bezeichnet (BMBF, 2007a, S. 11). Dabei werden mehrere flexible und kontextabhängige Teilprozesse unterschieden, die das Erkennen von Buchstaben und Wörtern und das Erfassen von Wortbedeutungen umfassen. Die reine Dekodierung geschriebener Informationen umfasst allerdings nicht das Integrieren von Sätzen zu Bedeutungseinheiten oder das Aufbauen einer kohärenten mentalen Repräsentation der Textinhalte. Diese Prozesse, die über die reine Dekodierungsleistung hinausgehen, werden als Textverstehen (oder, in Abgrenzung zum Hörverstehen, als Leseverständnis) bezeichnet.

2.3.2 Textverstehen

Zentraler Gegenstand der Textverstehensforschung sind die kognitiven Prozesse, die zur Konstruktion mentaler Repräsentationen von gesprochenen oder

[4] Eine Datenbanksuche in PsycINFO ergab zum Zeitpunkt der Verschriftlichung dieser Arbeit 9 508 Ergebnisse für den Suchbegriff “reading comprehension”, 4 744 Ergebnisse für “reading skills”, 1 123 Ergebnisse für “text comprehension” und 500 Ergebnisse für “reading literacy” (Abfrage vom 15. August 2011).

geschriebenen Texten führen, sowie die Variablen, die diese Prozesse beeinflussen. In frühen kognitionspsychologischen Ansätzen wurde Textverstehen dabei als ein Vorgang beschrieben, der im Wesentlichen durch das Aufbauen und Abrufen einer Repräsentation des Textes selbst gekennzeichnet ist. Die psychometrische Erfassung dieser Fähigkeit erfolgte dementsprechend mithilfe von Verfahren, die eine möglichst präzise, inhaltsgetreue oder gar wörtliche Wiedergabe von zuvor gelesenen Textpassagen testeten (Meyer, 1975). Diese Auffassung wurde jedoch zunehmend kritisch infrage gestellt und anhand verschiedener Experimente gezeigt, dass Personen Texte nicht zwangsläufig im Wortlaut, sondern vielmehr in Form situativer Modelle erinnern (Bransford, Barclay & Franks, 1972; Bransford & Johnson, 1972).

Ein früher theoretischer Ansatz zu dieser Form des Textverstehens, der sich vor allem auf das Hören von gesprochenen Texten konzentriert, in vielen Punkten aber auch auf die Verarbeitung schriftlicher Texte übertragen werden kann, stammt von van Dijk und Kintsch (1983). In diesem Ansatz wird der Vorgang der kognitiven Verarbeitung sprachlicher Informationen in mehrere Prozesse unterteilt, die einerseits nacheinander, andererseits auch parallel zueinander ablaufen und zum Teil als Bottom-Up-, zum Teil als Top-Down-Prozesse klassifiziert werden können.

Auf basaler Ebene müssen Geräusche zunächst als Phoneme (bzw. analog dazu: Schriftzeichen als Grapheme) wahrgenommen und Verknüpfungen aus diesen als Morpheme erkannt werden. Morpheme sind diejenigen sprachlichen Bestandteile, aus denen einzelne Wörter bestehen. Insofern es sich um geschriebene Texte handelt, sind diese basalen Fähigkeiten mit reiner Lesefähigkeit identisch.

Bestimmte Kombinationen aus Morphemen ergeben einfache Aussagen, die vom Rezipienten als solche identifiziert und zu möglichst kohärenten mentalen Abbildern des Textes verknüpft werden müssen. Die Interpretation des Textes ist dabei vom episodischen Gedächtnis bzw. vom individuellen Vor- und Weltwissen des Rezipienten abhängig, welches dazu selektiv aktiviert und mit den aus dem Text entnommenen Informationen integriert werden muss. Die auf diese Weise konstruierte individuelle mentale Repräsentation des Textes wird von van Dijk und Kintsch (1983) als situatives Modell bezeichnet. Das situative Modell ist “the representation of that fragment of the world the text

is speaking about" (S. 338), wobei im Text ein großer Teil dieses Ausschnittes aus der Realität gar nicht unmittelbar enthalten sein muss, sondern vom Rezipienten in Form einer Integration aus Textinhalt und Vorwissen konstruiert wird.

Die moderne Textverstehensforschung folgt dem von van Dijk und Kintsch vorgeschlagenen Ansatz und fasst den Prozess des Textverstehens als Konstruktion mentaler Repräsentationen auf (Kürschner & Schnotz, 2008; Mayer, 1997, 2005b; Schnotz, 2005, 2006). Die meisten Modelle übernehmen die Unterscheidung von zwei Formen mentaler Repräsentationen, wie sie bei van Dijk und Kintsch zu finden ist, verwenden aber zum Teil abweichende Bezeichnungen. Konzeptuelle mentale Abbilder eines Textes werden in der Regel als *propositionale Repräsentationen* bezeichnet. Propositionen bestehen "aus komplexen Symbolen, die mithilfe einfacher syntaktischer Regeln verbunden sind" (Kürschner & Schnotz, 2008, S. 140). Propositionale Repräsentationen sind also mentale Abbilder komplexer Symbolsysteme und ähneln in ihrer Struktur der Struktur des gelesenen Textes.

Bildhafte, situative oder modellartige kognitive Abbilder werden hingegen *mentale Modelle* genannt. Auch diese enthalten Informationen, die nicht explizit im Text auftauchen, sondern von den Lesenden durch Verknüpfung der Textinhalte mit individuellem Vorwissen implizit in den Prozess der Konstruktion der mentalen Repräsentation einbezogen werden. Mentale Modelle "repräsentieren Sachverhalte aufgrund inhärenter struktureller Eigenschaften entsprechend analoger Abbildungsprinzipien" (Kürschner & Schnotz 2008, S. 140).

Ein integriertes Modell des Text- und Bildverstehens.

Den bisher vorgestellten Erklärungsansätzen zum Verstehen von Texten ist gemein, dass sie jeweils auf bestimmte Darbietungs- und Wahrnehmungsformen beschränkt sind. Sie befassen sich entweder mit dem Verstehen gesprochener oder mit dem Verstehen geschriebener Texte, nicht aber mit der kognitiven Verarbeitung anderer Darbietungsformate. Sie sind somit auf die Dekodierung sprachlicher Informationen limitiert und grenzen dabei andere Textbestandteile wie Abbildungen, Tabellen oder Diagramme aus. Es ist jedoch anzunehmen, dass Personen auf Basis unterschiedlich dargebotener Informationen (z. B. auditiv vs. visuell oder textbasiert vs. bildhaft) zu ganz ähnlichen oder sogar

identischen mentalen Repräsentation gelangen können—und dass die kognitive Verarbeitung dieser Informationen dabei nicht völlig unterschiedlich abläuft. Um diesem Problem zu begegnen, schlägt Schnotz (2005) ein Modell vor, das die Bildung propositionaler Repräsentationen und mentaler Modelle auf Basis von vier unterschiedlichen Formen externer Repräsentationen beschreibt. Das Integrierte Modell des Text- und Bildverstehens (ITPC-Modell) berücksichtigt zwei Kodierungsformen (sprachliche und bildhafte Informationen) und zwei Wahrnehmungskanäle (auditiver und visueller Kanal). Daraus ergeben sich vier mögliche Kombinationen: gesprochener Text, geschriebener Text, grafische Abbildungen und Hörbilder (im englischen Originaltext: *Sound Images*). Die Besonderheit des ITPC-Ansatzes ist, dass er die Wahrnehmung und Verarbeitung von Informationen aus diesen vier möglichen Quellen in einem gemeinsamen Modell beschreibt (Abbildung 2.1).

Von den vier im Modell berücksichtigten Darbietungsmodi kommen zwei für die Anwendung in schriftlichen Tests infrage: geschriebene Texte und visuelle Abbildungen. Letztere haben sich bereits in zahlreichen Studien zum multimedialen Lernen als geeignet und hilfreich erwiesen, den Extraneous Cognitive Load—also die allein durch die Art der Informationsdarstellung verursachte kognitive Beanspruchung—bei der Informationsverarbeitung gegenüber Texten zu reduzieren und somit das Verstehen komplexer Sachverhalte zu erleichtern (Brünken, Seufert & Zander, 2005; Kozma, 2000). Derselbe Effekt wäre auch beim Testen wünschenswert—zumindest dann, wenn nicht die Fähigkeit des Lesens selbst gemessen werden soll, sondern andere Kompetenzen mithilfe textbasierter Aufgaben erfasst werden.

Sowohl schriftlich dargebotene Texte als auch Bilder werden im ITPC-Modell zunächst über denselben Wahrnehmungskanal wahrgenommen und verarbeitet (siehe rechte Seite in Abbildung 2.1). Der "Weg" der Informationen führt über das visuelle sensorische Register (Auge) zum visuellen Arbeitsgedächtnis. Auf der Arbeitsgedächtnisebene erfolgt dann eine Filterung und getrennte Weiterverarbeitung der Informationen. Stammen diese aus einem geschriebenen Text, erfolgt eine Verarbeitung auf dem verbalen Kanal. Unter Einbeziehung kognitiver Schemata aus dem Langzeitgedächtnis werden aus den entnommenen Textinhalten propositionale Repräsentationen gebildet, die im Laufe der weiteren Verarbeitung in mentale Modelle überführt werden können.

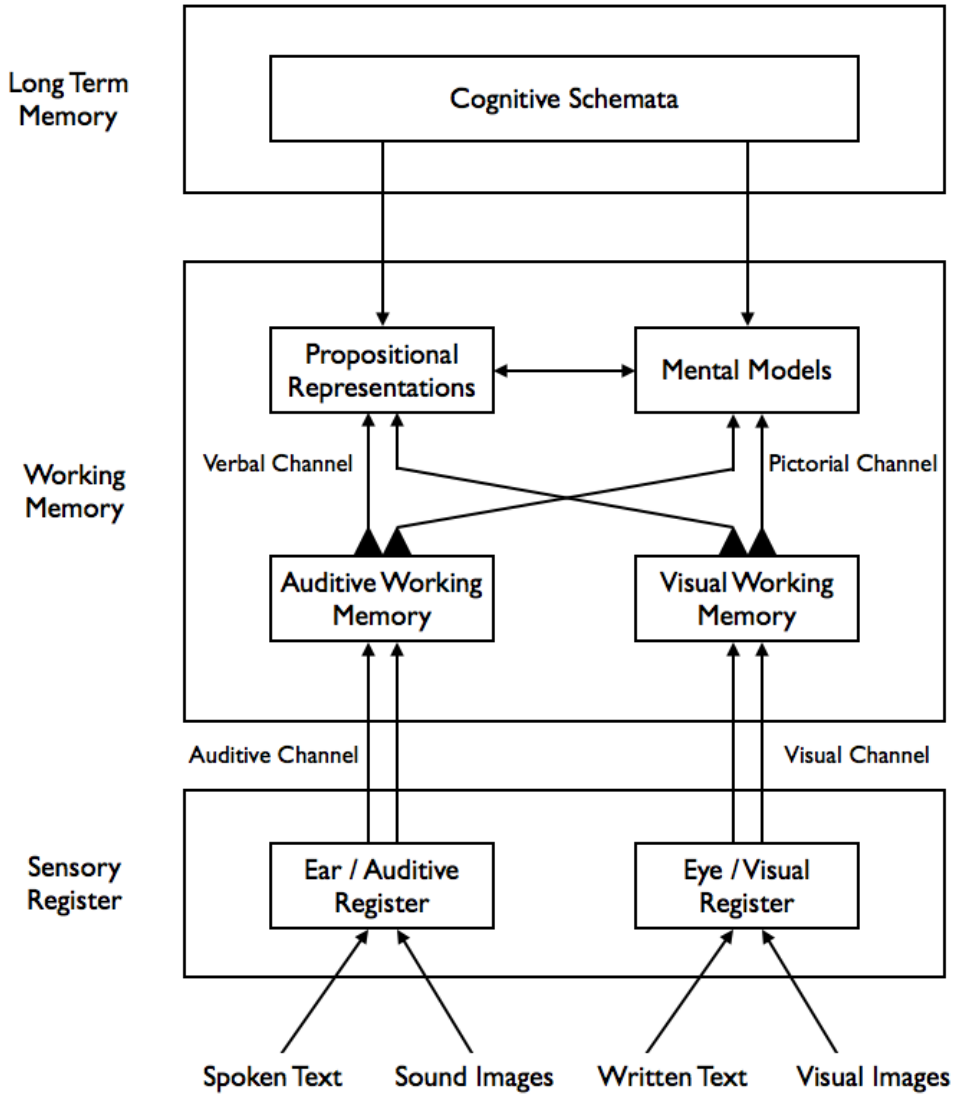


Abbildung 2.1: Das Integrierte Modell des Text- und Bildverstehens (eigene Darstellung nach Schnotz, 2005). Die Dreiecke (▲) oberhalb der Ebenen "Auditive Working Memory" und "Visual Working Memory" stellen Filtermechanismen dar.

Handelt es sich beim Ausgangsmaterial hingegen um Bilder, erfolgt eine Weiterverarbeitung auf dem piktorialen Kanal. Dabei führt das Integrieren von Informationen aus dem Langzeitgedächtnis zur Bildung mentaler Modelle, aus denen wiederum propositionale Repräsentationen abgeleitet werden können.

Analog dazu werden gesprochene Texte und Hörbilder zunächst über das auditive Register (die Hörorgane) wahrgenommen (vgl. linke Seite in Abbildung 2.1). Sie gelangen über den auditiven Kanal ins auditive Arbeitsgedächtnis. Dort erfolgt wiederum eine Filterung und Aufteilung: Gesprochene Texte werden auf dem verbalen Kanal weiterverarbeitet und führen unter Einbeziehung kognitiver Schemata aus dem Langzeitgedächtnis zunächst zur Bildung propositionaler Repräsentationen. Hörbilder werden hingegen auf dem piktorialen Kanal weiterverarbeitet und durch das Integrieren von Informationen aus dem Langzeitgedächtnis in mentale Modelle überführt. Im weiteren Verlauf der Informationsverarbeitung können auch hier aus propositionalen Repräsentationen mentale Modelle gebildet bzw. aus mentalen Modellen propositionale Repräsentationen abgeleitet werden.

Das ITPC-Modell statuiert, dass sowohl die sensorischen Kanäle (auditiver und visueller Kanal) als auch die Elemente im Bereich des auditiven und visuellen Arbeitsgedächtnisses in ihrer Verarbeitungs- und Speicherkapazität beschränkt sind. Aus der zentralen Annahme getrennter Wahrnehmungs- und Verarbeitungskanäle wird unter anderem darauf geschlossen, dass Verstehens- und Lerneffekte besser ausfallen, wenn statt reinem Text (verbaler Kanal) eine Kombination aus Text und Bildern (verbaler und piktorialer Kanal) als Informationsquelle vorliegt. Dieser Effekt, der von Mayer (1997) als *Multimedia-Effekt* bezeichnet wurde, hat sich in zahlreichen Studien empirisch bewährt (Mayer, 2005b). Voraussetzungen für eine effiziente Verarbeitung von Texten und zugehörigen Bildern sind eine semantische Nähe (Kohärenz) sowie eine räumliche und zeitliche Nähe (Kontiguität) der dargebotenen Informationen (Mayer, 2003).

Die Kombination aus Text und Bildern bietet aber noch einen weiteren Vorteil beim Verstehen und Lernen: Ist die Informationsaufnahme oder -verarbeitung auf einem Kanal gestört oder beeinträchtigt, so kann der damit einhergehende Verlust an Informationen über einen anderen Kanal ausgeglichen werden, wenn den Rezipienten entsprechende externe Informationen

vorliegen (Schnotz, 2005, S. 62). Daraus folgt unter Anderem, dass schwache Leserinnen und Leser in höherem Maß vom Einsatz von Abbildungen in Texten profitieren können als Personen mit guten Leseleistungen. Analog dazu profitieren auch Personen mit niedrigem Vorwissen stärker von Abbildungen in Texten als Personen mit hohem Vorwissen (ebd.).

Das Integrierte Modell des Text- und Bildverstehens und die daraus abgeleiteten Schlussfolgerungen hinsichtlich des Verstehens von Texten ohne und mit Bildern stellen das theoretische Gerüst für die in dieser Dissertation aufgestellten Hypothesen über die Effekte unterschiedlicher Darbietungsformate in Testaufgaben dar. Da im Rahmen dieser Arbeit nur geschriebene, jedoch keine gesprochenen Texte zur Anwendung kommen, soll—in Abgrenzung zum Hörverstehen—statt des Begriffs *Textverstehen* die Bezeichnung *Leseverständnis* verwendet werden.

2.3.3 Lesekompetenz

In den Abschnitten 2.3.1 und 2.3.2 wurde Lesen als die Fähigkeit zum Dekodieren schriftlicher Informationen definiert und auf den Begriff des Leseverständnisses im Sinne einer Konstruktion mentaler Repräsentationen geschriebener Texte ausgeweitet. In der pädagogisch-psychologischen Diagnostik kommen ganz unterschiedliche Verfahren zum Einsatz, um die Fähigkeit des Verstehens von Texten zu beschreiben und zu messen. In Large-Scale-Assessments wird hierzu häufig der bereits vorgestellte Kompetenzbegriff von Weinert herangezogen und von *Lesekompetenz* gesprochen. Innerhalb der PISA-Studien ist Lesekompetenz als die Fähigkeit definiert, "geschriebene Texte zu verstehen, zu nutzen und über sie zu reflektieren, um eigene Ziele zu erreichen, das eigene Wissen und Potenzial weiterzuentwickeln und am gesellschaftlichen Leben teilzunehmen" (Prenzel et al., 2004, S. 20). Lesekompetenz geht über das Konzept reiner Dekodierung (Lesefähigkeit) sowie über die Konstruktion mentaler Repräsentationen (Leseverständnis) hinaus und erweitert diese um die Fähigkeit, "in Alltagssituationen aktiv und konstruktiv mit geschriebenen Texten umzugehen" (Artelt, Drechsel, Bos & Stubbe, 2008, S. 36). Sie ist nicht auf Text als ausschließliches Darbietungsformat reduziert (Baumert, Brunner, Lüdtke & Trautwein, 2007), sondern bezieht sich auf schriftliche Dokumente, "in denen sowohl verbale Informationen in Form von Schriftzeichen (graphemisch)

als auch piktoriale Informationen in Form von Bildzeichen (graphisch) enthalten sind" (Schnotz & Dutke, 2004, S. 63). So gesehen ist im Konstrukt der Lesekompetenz sowohl die Fähigkeit des Textverstehens als auch die Fähigkeit zum Verstehen von visuellen Abbildungen enthalten.

Um die drei Begriffe klar voneinander abzugrenzen, sollen Lesefähigkeit, Leseverständnis und Lesekompetenz im Rahmen dieser Arbeit folgendermaßen definiert sein:

- *Lesefähigkeit:* die Fähigkeit, geschriebene Buchstaben, Wörter und Texte zu dekodieren,

- *Leseverständnis:* die Fähigkeit zur Konstruktion kohärenter mentaler Repräsentationen auf Basis geschriebener Texte unter Einbeziehung des individuellen Vor- und Weltwissens,

- *Lesekompetenz:* die Fähigkeit, multiple Darstellungen in schriftlichen Dokumenten zu verstehen, sie aktiv zu nutzen und über sie zu reflektieren, sowie sie zum Erreichen von Zielen, zur Weiterentwicklung von Wissen und Potenzialen und zur gesellschaftlichen Teilhabe anzuwenden.

2.3.4 Lesegeschwindigkeit

Neben der Genauigkeit der Dekodierung und dem individuellen Vor- und Weltwissen ist die Lesegeschwindigkeit ein weiterer Indikator für die erfolgreiche Konstruktion einer mentalen Repräsentation eines Textes (Artelt et al., 2007; Landerl, 2001). Die Lesegeschwindigkeit hängt in erheblichem Maß mit der Effizienz beim Zugriff auf das semantische Lexikon zusammen und ist für die Qualität des Lesevorgangs essentiell (Perfetti, 1985; Perfetti & Hogaboam, 1975).

Verschiedene Experimente liefern Hinweise dafür, dass sich starke und schwache Leserinnen und Leser nicht zwangsläufig in der Exaktheit der Dekodierung einzelner Wörter, wohl aber hinsichtlich der dafür benötigten Zeit unterscheiden: "if a given sample of words does not produce differences in decoding accuracy [between high-ability and low-ability readers], it will in speed" (Perfetti, 1985, S. 91). Der Geschwindigkeitsnachteil leseschwacher Personen

wird von Perfetti auf einen erhöhten kognitiven Verarbeitungsaufwand zurückgeführt und nimmt mit der Anzahl der zu dekodierenden Silben bzw. mit der Anzahl der zu dekodierenden Wörter zu. Beim Dekodieren von Abbildungen, Zahlenfolgen oder Farben zeigten sich hingegen keine statistisch bedeutsamen Geschwindigkeitsunterschiede zwischen starken und schwachen Leserinnen und Lesern.

Perfetti führt weiter aus, dass Defizite auf der Ebene der Worterkennung beim Lesen längerer oder schwieriger Texte kognitive Ressourcen beanspruchen, die eigentlich für hierarchiehöhere Verstehensprozesse benötigt würden: "Efficient lexical access, rapid and low in resource cost, enables working memory to carry out the propositional text work. Inefficient lexical access, slow and effortful, makes it more difficult for working memory to do this work" (Perfetti, 1985, S. 113). Langsamen Leserinnen und Lesern fehlt demnach die nötige Verarbeitungskapazität für die Bildung mentaler Repräsentationen des gelesenen Textes, wodurch sich "die Kohärenzbildung auf der Ebene des Textes erschwert" (Artelt et al., 2007, S. 50). Auch Schneider, Schlagmüller und Ennemoser (2007) folgen diesem Ansatz und betonen, dass die Möglichkeiten einer aktiven Sinnentnahme deutlich eingeschränkt sind, wenn das Dekodieren der Textinformationen ein zu hohes Maß an Zeit erfordert.

Eine niedrige Lesegeschwindigkeit ist also vermutlich die Folge von mangelnder Effizienz bzw. eines zu hohen kognitiven Aufwands bei Worterkennungsprozessen sowie gleichzeitig die Ursache für Verstehensprobleme auf höheren Verarbeitungsebenen. Bezogen auf Testsituationen, in denen schriftliche Aufgaben unter Zeitdruck zu lösen sind, müsste eine niedrige Lesegeschwindigkeit zwangsläufig einen erheblichen Einfluss auf die Anzahl richtig gelöster Aufgaben haben. Aber auch in Tests, bei denen eine ausreichend lange Bearbeitungszeit gewährleistet ist, ist anzunehmen, dass langsame Leserinnen und Leser aufgrund der eingeschränkten Sinnentnahme und der höheren Beanspruchung kognitiver Ressourcen stark benachteiligt sind. Dies müsste insbesondere für Personen gelten, die mehrsprachig aufgewachsen sind, da deren Lesegeschwindigkeit durch Probleme beim Prozess der Worterkennung gegenüber Muttersprachlern besonders niedrig ausfällt (Martiniello, 2009; Stanat & Schneider, 2004).

2.4 Zusammenhänge zwischen Leseleistungen und naturwissenschaftlichen Kompetenzen in schriftlichen Tests

Wie im vorangegangenen Kapitel beschrieben, wird unter Leseverständnis die Konstruktion kohärenter mentaler Modelle auf Basis geschriebener Texte unter Einbeziehung des individuellen Vor- bzw. Weltwissens verstanden. Genau diese Fähigkeit ist auch zum Lösen einer beliebigen schriftlichen Testaufgabe erforderlich—ganz gleich, ob sie Leistungen in Biologie, Mathematik oder Betriebswirtschaftslehre messen soll. Nur wer in der Lage ist, angemessene mentale Repräsentationen des Aufgabenstimulus' und der Fragestellung zu konstruieren, kann unter Aktivierung von vorhandenem Vorwissen bzw. fachlicher oder methodischer Kompetenz überhaupt zur richtigen Lösung gelangen. Es sind letztere Fähigkeiten, die eine schriftliche Aufgabe in einem bestimmten Fach von einer reinen Leseverständnis-Aufgabe unterscheiden. Ist eine Person allerdings nicht in der Lage, einen Aufgabentext richtig zu verstehen bzw. zu interpretieren, wird sie die Aufgabe trotz vorhandener fachlicher oder methodischer Kompetenz nicht (bzw. mit stark reduzierter Wahrscheinlichkeit) lösen können—das Instrument ist mit Leseverständnis konfundiert: "In other words, a test cannot provide valid information about a student's knowledge or skills if a language barrier prevents the students from demonstrating what they know and can do" (National Research Council, 2000, S. 20).

Textbasierten Testaufgaben ist also gemein, dass ihre Bearbeitung zwangsläufig das Lesen und Verstehen schriftlich dargebotener Informationen erfordert. Wenn—wie im Fall der meisten Aufgaben zur Messung von Kompetenzen—bestimmte lösungsrelevante Informationen bereits im Aufgabenstimulus enthalten sind und die Fähigkeit getestet wird, diese Informationen richtig zu entnehmen und auf alltagsnahe Situationen anzuwenden, dann wird dies gelegentlich zum Anlass genommen, die Validität der entsprechenden Instrumente generell infrage zu stellen. Beispielsweise kritisiert Klein (2010), dass im Fall kompetenzorientierter Aufgaben aus dem nordrhein-westfälischen Zentralabitur im Fach Biologie Lesekompetenz ausreiche, um zu einer richtigen Lösung zu gelangen, während grundlegendes biologisches Fachwissen nicht

erforderlich sei (S. 15). Rindermann (2006) argumentiert, dass die PISA-Naturwissenschaftsaufgaben das Lesen längerer Texte voraussetzen, und dass Teile dieser Aufgaben "auch unter Lesefähigkeit subsumierbar" seien (S. 73). Aber auch Aufgaben, die weniger Leseaufwand erfordern, seien häufig allein "durch Informationsentnahme aus den gegebenen Materialien" lösbar und wiesen "große Überlappungen ... zu den Leseaufgaben" auf (S. 74). Diese Überlappungen spiegeln sich ganz offensichtlich auch in den hohen Korrelationen wider, die bei verschiedenen Analysen der PISA-Daten zwischen den Skalen *Lesekompetenz* und *Naturwissenschaften* gefunden wurden. Für PISA 2003 wird hierzu beispielsweise eine latente Korrelation von .86 berichtet (OECD, 2003, S. 36). In der deutschen Teilstichprobe liegt der Zusammenhang bei .87 (Leutner et al., 2004, S. 167). Wird die kognitive Grundfähigkeit kontrolliert, beträgt der Wert immer noch .77 (ebd.).

Wenngleich Rindermanns Kritik teilweise entkräftet werden konnte (Baumert et al., 2007; Prenzel, Walter & Frey, 2007), so können die hohen Korrelationen dennoch als Indiz dafür gewertet werden, dass Naturwissenschafts- und Leseaufgaben eine gewisse Nähe zueinander aufweisen. Die Frage, ob diese Nähe tatsächlich auf eine hohe Textlastigkeit der Aufgaben zurückgeht oder ob sie mit der Rolle des Lesens beim Kompetenzerwerb erklärbar ist, lässt sich hingegen mit diesen Korrelationen allein nicht beantworten.

Ein Hinweis darauf, dass die Ergebnisse von Tests zur Erfassung naturwissenschaftlicher Kompetenzen mit Leseverständnis konfundiert sind, findet sich bei Prenzel et al. (2002). In der Studie wurden Regressionsanalysen mit Items aus dem PISA-Naturwissenschaftstest durchgeführt. Dabei gingen verschiedene Aufgabenmerkmale als Prädiktoren für die Lösungswahrscheinlichkeit in die Analysen ein. Die Autoren stellten fest, "dass die Vorgabe einer Grafik bzw. eines stilisierten Bildes ... die Lösung der Aufgabe erleichtert" (S. 132). Dies wird auf eine Reduzierung des zum Lösen der Aufgabe erforderlichen Textverständnisses zurückgeführt. Martiniello (2009) berichtet für Schülerinnen und Schüler aus fremdsprachigen Elternhäusern, dass sich bildhafte Repräsentationen in Testitems eignen, um den Einfluss mangelnder sprachlicher Fähigkeiten auf die Lösung von Mathematik-Aufgaben zu reduzieren. Allerdings wird auch darauf verwiesen, dass die Generalisierbarkeit solcher Befunde limitiert ist, weil der Effekt u. a. stark von der Art der verwendeten Abbildungen abhängt. So sei

es auch vorstellbar, "die Aufgabenschwierigkeit zu erhöhen, wenn etwa besonders knifflige Grafiken zur Lösung einer Aufgabe herangezogen werden müssen" (Prenzel et al., 2002, S. 133). Ein weiterer Grund für eine mangelnde Verallgemeinerbarkeit sind quasi-experimentelle Versuchsanordnungen, in denen das Vorhandensein von Abbildungen zudem nur eines von vielen Merkmalen ist, in denen sich die untersuchten Aufgaben unterscheiden. Um eine zuverlässige Aussage über den Einfluss von Grafiken auf die Lösungswahrscheinlichkeit treffen zu können, müsste man streng genommen alle anderen Aufgabenmerkmale konstant halten und ausschließlich das Darbietungsformat der lösungsrelevanten Informationen (in den Ausprägungen *Text* vs. *Bild*) verändern.

2.4.1 Maßnahmen im Projekt ESNaS

Basierend auf den Erkenntnissen aus PISA und anderen Erhebungen wurde bei der Konstruktion von Aufgaben für das Instrument zur Evaluation der Bildungsstandards eine Reihe von Kriterien eingeführt, um den Leseaufwand und eine damit einhergehende Konfundierung mit Leseverständnis von vorneherein auf einem möglichst niedrigen Niveau zu halten (Kauertz et al., 2010). Die Richtlinien für die sprachliche Gestaltung der Aufgaben orientierten sich unter anderem an Befunden von Sumfleth und Schüttler (1995), die positive Effekte bestimmter Textgestaltungskriterien beim Lernen mit naturwissenschaftlichen Texten nachweisen konnten. So wurden die Aufgabenentwicklerinnen und Aufgabenentwickler des ESNaS-Projektes beispielsweise instruiert, die Aufgabentexte möglichst auf unmittelbar zum Lösen notwendige Inhalte zu beschränken und auf schmückende, redundante oder unwesentliche Elemente zu verzichten. Im Rahmen der Bewertungs- und Korrekturschleifen, die jede Aufgabe zu durchlaufen hatte, wurden die Aufgabentexte zudem durch das Institut für Sprache und Didaktik II der Universität zu Köln überprüft und bei Bedarf so überarbeitet, dass auch schwächere Leserinnen und Leser in der zu testenden Altersgruppe in der Lage sein sollten, die Texte zu verstehen. Um eine mögliche Konfundierung mit Leseleistungen weiter zu reduzieren, wurden im Fach Biologie die lösungsrelevanten Informationen in den Aufgaben immer dann in Form von Abbildungen dargestellt, wenn dies praktisch umsetzbar erschien. Die Abbildungen wurden als realistische Graustufenzeichnungen von einer professionellen Grafikerin in enger Zusammenarbeit mit der fachdidak-

tischen Leitung angefertigt, wobei neben der fachlichen Richtigkeit auch auf eine höchstmögliche Einfachheit und Eindeutigkeit der Darstellungen geachtet wurde. Daneben wurden teilweise auch Bilder eingesetzt, die keine unmittelbar für das Lösen der Aufgaben relevanten Informationen enthielten. Dabei handelte es sich in erster Linie um fotografische oder zeichnerische Darstellungen von Pflanzen oder Tieren, bei denen man nicht sicher sein konnte, dass die Schülerinnen und Schüler diese allein anhand des Namens identifizieren können (z. B. Nördliches Gleithörnchen, Nagel-Manati oder Pferdespringer).

Neben den genannten Maßnahmen, die der Reduktion einer möglichen Konfundierung mit Leseverständnis dienen sollen, ergeben sich durch die systematische Kontrolle schwierigkeitserzeugender Aufgabenmerkmale allerdings auch strukturelle Gemeinsamkeiten der ESNaS-Testitems mit typischen Leseverständnis-Aufgaben. Insbesondere im unteren Schwierigkeitsbereich ist zum erfolgreichen Lösen der Aufgaben nur das Reproduzieren oder Selegieren vorgegebener Informationen erforderlich. Hiervon ist ein großer Teil der Aufgaben betroffen (vgl. hierzu Kap. 2.2). Betrachtet man gängige Verfahren zur Erfassung von Leseverständnis oder Lesekompetenz (Hemminger, Roth, Schneck, Jans & Warnke, 2000), so fällt auf, dass auch diese mitunter die Fähigkeit zur selektiven Informationsentnahme erfassen. Insofern besteht eine konzeptuelle Ähnlichkeit zwischen den ESNaS-Aufgaben der unteren Schwierigkeitsbereiche und Leseverständnis-Aufgaben. Die fachdidaktischen Verantwortlichen des ESNaS-Projektes gehen selbst davon aus, dass die Aufgaben zum Reproduzieren und Selegieren "stark mit Textverstehen konfundiert" sind (Kauertz et al., 2010, S. 145). Auch in den höheren Schwierigkeitsbereichen bildet die Entnahme von Informationen aus dem Aufgabenstimulus eine wichtige Voraussetzung für das erfolgreiche Lösen (ebd.). Es stellt sich somit die Frage, "wie hoch der Einfluss ... der Lesekompetenz auf die so operationalisierte naturwissenschaftliche Kompetenz" generell ist (Kauertz et al., S. 150).

Für das ESNaS-Instrument sind bisher keine Korrelationen zwischen der Testleistung und Lesekompetenz publiziert. Im Rahmen einer Pilotierungsstudie aus dem Jahr 2009 wurden projektintern allerdings erste Berechnungen vorgenommen, die zumindest eine tendenzielle Einschätzung der Zusammenhänge erlauben. Auf Basis einer nicht repräsentativen Teilstichprobe ($N = 291$) ergab sich dabei zwischen den Leistungen in Biologie (Kompetenzbereich *Erkenntnis-*

gewinnung, 76 Items) und einem Lesekompetenztest des IQB (27 Items) eine latente Korrelation von .58 (eigene Berechnungen in ConQuest). Wenngleich ein direkter Vergleich mit den entsprechenden Kennwerten aus PISA aufgrund der Unterschiedlichkeit der Instrumente nicht sinnvoll ist, könnte man dieses Ergebnis dennoch als Indiz dafür heranziehen, dass auch die im Rahmen der ESNaS-Pilotierungsstudie erhobenen Leistungen zu einem gewissen Maß mit Lesekompetenz konfundiert sind. Diese Interpretation entspricht den Erwartungen der fachdidaktischen Leitungen.

2.4.2 Mögliche Moderatoreffekte

In Kapitel 2.3 wurde bereits darauf hingewiesen, dass für bestimmte Personengruppen (z. B. langsame Leserinnen und Leser, Personen mit geringem Vorwissen und mehr- oder fremdsprachig aufgewachsene Personen) Besonderheiten beim Lesen von Texten gelten. Wie im Folgenden gezeigt wird, wirken sich diese Besonderheiten u. a. auf den Erfolg beim Lernen mit Texten und Bildern aus. Derartige Interaktionen aus der Wirksamkeit unterschiedlicher Methoden (*Treatments*) in Abhängigkeit von Voraussetzungen (*Aptitudes*) auf Seiten der Personen werden unter dem Begriff *Aptitude-Treatment-Interaction* oder kurz ATI zusammengefasst (Cronbach & Snow, 1969). In der pädagogischen Psychologie haben sich ATI-Analysen als geeignet erwiesen, um "individuelle Unterschiede beim Lernen mit neuen Medien" zu untersuchen (Brünken & Leutner, 2005, S. 25). Auf Basis der Annahmen und Befunde zum Lernen mit Texten und Bildern werden im folgenden Abschnitt mögliche Moderatorvariablen identifiziert, von denen angenommen wird, dass sie auch beim Testen mit Text- und Text-Bild-Aufgaben eine Rolle spielen. Die entsprechenden Effekte sollen im Rahmen der Vorstudien und der Haupterhebung zusätzlich untersucht werden.

Im Integrierten Modell des Text- und Bildverstehens (Schnotz, 2005) wird davon ausgegangen, dass Einschränkungen auf einem Wahrnehmungs- und Verarbeitungskanal ausgeglichen werden können, wenn alternative Informationsquellen zur Verfügung stehen, deren Wahrnehmung und Verarbeitung auf anderen Kanälen stattfinden. Dementsprechend müssten Personen, deren Leseleistungen schwach sind, stärker vom Einsatz von Abbildungen profitieren, als Personen mit guten Leseleistungen (Schnotz, 2005, S. 62). Die von Schnotz

publizierten Annahmen und Befunde hierzu beziehen sich allerdings vorrangig auf das Lernen mit Texten und Bildern, nicht auf die Rolle unterschiedlicher Formate beim Testen mit schriftlichen Aufgaben. Da der Verstehensprozess aber auch für das Bearbeiten von Testaufgaben elementar ist, kann davon ausgegangen werden, dass die beschriebenen Effekte gleichermaßen für Situationen gelten, in denen Kompetenzen mit schriftlichen Instrumenten gemessen werden sollen.

Des Weiteren postuliert das ITPC-Modell, dass Defizite im individuellen Vor- und Weltwissen durch den Einsatz von Abbildungen teilweise ausgeglichen werden können. Aus diesem Grund profitieren Personen mit niedrigem Vor- und Weltwissen beim Lernen in höherem Maß vom Einsatz von Abbildungen als Personen mit hohem Vorwissen (Schnotz, 2005, S. 62). Auch diese Annahme wird im Rahmen der vorliegenden Arbeit auf die Situation des Testens mit Papier-und-Bleistift-Aufgaben ausgeweitet.

Stanat und Schneider (2004) weisen darauf hin, dass mehrsprachig aufgewachsene Personen über einen geringeren Wortschatz pro Sprache verfügen als Personen, die einsprachig aufgewachsen sind. Aus diesem Grund benötigen sie mehr Zeit zum Dekodieren, was sich in einer niedrigeren Lesegeschwindigkeit äußert. Hierunter leidet die Sinnentnahme, d. h. die Konstruktion einer kohärenten mentalen Repräsentation des Textes (Landerl, 2001; Schneider et al., 2007). Davon ist sowohl die Erstsprache betroffen als auch die zweite und ggf. weitere Sprachen (Stanat & Schneider, 2004, S. 251). Da Abbildungen sprachunabhängig sind und Beeinträchtigungen bei der Wahrnehmung und Verarbeitung verbaler Informationen kompensieren können (Schnotz, 2005, S. 62), müssten fremd- oder mehrsprachig aufgewachsene Personen in höherem Maß von Abbildungen profitieren als Personen, in deren Elternhaus ausschließlich Deutsch gesprochen wird.

2.5 Hypothesen

Basierend auf den Annahmen des Integrierten Modells des Text- und Bildverstehens (Schnotz, 2005) und den Befunden von Prenzel et al. (2002) ist anzunehmen, dass der Multimedia-Effekt (Mayer, 1997) nicht nur für das Lernen mit Texten und Bildern gilt, sondern auch für die Verstehensleistungen beim Bearbeiten von Testaufgaben—und somit für deren Lösungswahrscheinlichkeit. Defizite beim Verstehen einer Aufgabe können dazu führen, dass sie falsch gelöst wird, obwohl die Person über ausreichende fachliche oder methodische Kompetenzen verfügt (Martiniello, 2009; National Research Council, 2000). Je besser eine Person den Aufgabenstimulus und die Fragestellung versteht, umso höher ist die Wahrscheinlichkeit, dass sie die Aufgabe unter Aktivierung vorhandener Kompetenz richtig löst. Da Texte mit Bildern bessere Verstehensleistungen erzielen als Texte ohne Bilder, lautet die erste Hypothese:

H1: Mit kombinierten Text-Bild-Aufgaben werden bessere Leistungen erzielt als mit reinen Textaufgaben.

Das Integrierte Modell des Text- und Bildverstehens (Schnotz, 2005) postuliert, dass Einschränkungen auf einem Wahrnehmungs- und Verarbeitungskanal kompensiert werden können, wenn Informationen in einem alternativen Format angeboten werden, dessen Verarbeitung auf einem anderen Kanal erfolgt. Martiniello (2009) empfiehlt Abbildungen als alternatives Darbietungsformat für Personen mit eingeschränkten sprachlichen Fähigkeiten. Beim multimedialen Lernen profitieren Personen, deren Leseverständnis schwach ist, im Vergleich zu Personen mit hohem Leseverständnis besonders vom Einsatz von Abbildungen (ATI-Effekt). Dieser Effekt müsste auch beim Testen mit textbasierten Aufgaben zu finden sein. Die zweite Hypothese lautet dementsprechend:

H2: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Leseverständnis und Darbietungsformat feststellbar.

Neben dem Leseverständnis ist die Lesegeschwindigkeit ein zweiter wichtiger Indikator für die Fähigkeit des sinnerfassenden Lesens (Landerl, 2001; Schneider et al., 2007; Stanat & Schneider, 2004). Eine geringe Lesegeschwindigkeit ist sowohl eine Folge von Problemen bei der Dekodierung als auch eine Ursache für Einschränkungen auf höheren Verarbeitungsebenen (Perfetti, 1985; Perfetti & Hogaboam, 1975). In Analogie zu Hypothese (H2) lautet deshalb die dritte Hypothese:

H3: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Lesegeschwindigkeit und Darbietungsformat feststellbar.

Defizite im individuellen Vorwissen können beim Lernen mit Texten durch den Einsatz von Abbildungen teilweise ausgeglichen werden (Schnotz, 2005). Insofern müssten auch beim Testen vor allem Schülerinnen und Schüler mit geringerem Vorwissen von lösungsrelevanten Abbildungen profitieren. Die vierte Hypothese lautet deshalb:

H4: Personen mit niedrigem Vorwissen profitieren in höherem Maß vom Einsatz von Abbildungen als Personen mit hohem Vorwissen.

Martiniello (2009) sowie Stanat und Schneider (2004) weisen auf spezifische Defizite mehrsprachig aufgewachsener Personen hin. Der kleinere Wortschatz dieser Personen geht mit höherem Aufwand bei der Dekodierung und langsamerer Lesegeschwindigkeit einher. Da Abbildungen im Gegensatz zu Text nicht an Sprache oder Wortschatz gebunden sind, ergibt sich die fünfte Hypothese:

H5: Schülerinnen und Schüler, die fremd- oder mehrsprachig aufwachsen, profitieren in höherem Maß vom Einsatz von Abbildungen als solche, in deren Elternhaus ausschließlich Deutsch gesprochen wird.

Kapitel 3

Analyse der ESNaS-Pilotierungsdaten

Auf Basis der vorgestellten Theorien und Befunde wurde im vorangegangenen Kapitel die Hypothese aufgestellt, dass Aufgaben mit lösungsrelevanten Abbildungen leichter zu lösen sind als reine Textaufgaben. Dementsprechend müssten sich in den Daten des Projektes ESNaS Zusammenhänge zwischen dem Darbietungsformat der lösungsrelevanten Informationen (*Text* vs. *Abbildungen*) und der Lösungswahrscheinlichkeit der Aufgaben finden lassen. Dabei "kann die Erklärung von IRT-basierten Itemschwierigkeiten durch Aufgabenmerkmale ... als eine Prüfung der Konstruktvalidität betrachtet werden" (Moosbrugger, 2008a, S. 250). So ist ein Einfluss von Merkmalen, die Bestandteil des Kompetenzmodells sind, im Sinne einer konvergenten Validierung als Beleg für die Validität des Tests zu werten. Einflüsse von Aufgabenmerkmalen, die nicht Bestandteil des Kompetenzmodells sind, sind im Sinne einer diskriminanten Validierung hingegen als Hinweise auf eine mögliche Konfundierung zu betrachten.

Zum Zeitpunkt der Erstellung dieser Arbeit sind noch keine entsprechenden Befunde zu den Aufgaben zur Evaluation der Bildungsstandards in Biologie publiziert. Auf Basis von Daten aus der ESNaS-Pilotierungsstudie sollen deshalb zunächst Berechnungen angefertigt werden, die empirisch gestützte Aussagen zu dem prognostizierten Zusammenhang erlauben. Die Befunde werden dann als Grundlage für die in Kapitel 4 vorgestellte experimentelle Untersuchung dienen.

3.1 Ausgangssituation und Ziele

Bevor die Bildungsstandards in den Naturwissenschaften bundesweit evaluiert werden können, bedarf es einer Reihe von Voruntersuchungen, die der Validierung und Optimierung der eingesetzten Instrumente dienen. Für die Kompetenzbereiche *Umgang mit Fachwissen* und *Erkenntnisgewinnung* wurde begleitend zur Aufgabenkonstruktion in den Jahren 2008 und 2009 in jedem der drei Fächer Biologie, Chemie und Physik eine Reihe unabhängiger Präpilotierungen durchgeführt. Anschließend kamen alle konstruierten Aufgaben in einer gemeinsamen, fächerübergreifenden Pilotierung zum Einsatz. Diese Pilotierung wurde auch genutzt, um verschiedene zusätzliche Fragestellungen in projektbegleitenden Dissertationen zu untersuchen—im Fach Biologie beispielsweise von Kampa (vgl. Kampa, 2010) und Wellnitz (2012).

In der hier vorliegenden Arbeit sollen mithilfe der Pilotierungsdaten Antworten auf die Frage gefunden werden, ob die in Kapitel 2 diskutierten Aufgabenmerkmale einen Einfluss auf die Lösungswahrscheinlichkeit der Biologie-Items haben. Dabei interessiert insbesondere, inwieweit sich der von Prenzel et al. (2002) gefundene Effekt reproduzieren lässt, dass Naturwissenschafts-Aufgaben mit Abbildungen leichter zu lösen sind als Aufgaben ohne Bilder. Gemäß den in Kapitel 2 vorgestellten theoretischen Grundlagen werden entsprechende Berechnungen für die ESNaS-Aufgaben im Fach Biologie (Kompetenzbereich *Erkenntnisgewinnung*) durchgeführt.

3.2 Zu prüfende psychologische Hypothesen

Der Multimedia-Effekt (Mayer, 1997) gilt vermutlich nicht nur für das Lernen mit Texten und Bildern, sondern auch für das Bearbeiten von Testaufgaben. Unter Berücksichtigung der Annahmen des Integrierten Modells des Text- und Bildverstehens (Schnotz, 2005) und der Befunde von Prenzel et al. (2002) ist anzunehmen, dass die Lösungswahrscheinlichkeit von Testitems höher ausfällt, wenn lösungsrelevante Abbildungen in den Aufgabenstimuli enthalten sind. Die erste Hypothese lautet deshalb:

$H1_{Pilotierung}$: Der von Prenzel et al. (2002) gefundene Effekt, dass Naturwissenschafts-Aufgaben mit Abbildungen leichter zu lösen sind als Aufgaben ohne Bilder, lässt sich mit den ESNaS-Items zum Kompetenzbereich *Erkenntnisgewinnung* im Fach Biologie replizieren.

Daneben sollen mit der Komplexität und dem kognitiven Prozess zwei Aufgabeneigenschaften in die Berechnungen eingehen, von denen—gemäß dem in Kapitel 2.3.2 beschriebenen Kompetenzmodell—ebenfalls ein Einfluss auf die Itemschwierigkeit bekannt ist (Wellnitz, 2012). Dementsprechend lauten die zu prüfenden Hypothesen:

$H2_{Pilotierung}$: Die Komplexität der zum Lösen zu verarbeitenden Informationen hat einen Einfluss auf die Itemschwierigkeit.

$H3_{Pilotierung}$: Der zum Lösen der Aufgabe erforderliche kognitive Prozess hat einen Einfluss auf die Itemschwierigkeit.

3.3 Methode

Zur empirischen Überprüfung der Frage, welchen Einfluss bestimmte Aufgabenmerkmale auf die Itemschwierigkeit haben, bieten sich verschiedene Methoden an (vgl. Moosbrugger, 2008a, S. 249). Um den von Prenzel et al. (2002) gefundenen Effekt, dass Aufgaben mit Abbildungen leichter zu lösen sind als Aufgaben ohne Abbildungen, zu replizieren, ist das dort eingesetzte Verfahren einer linearen Regressionsanalyse auch in der hier vorliegenden Studie zur Anwendung gekommen. Dabei gehen mehrere Aufgabenmerkmale als Prädiktoren und die Itemparameter (Logits) als Kriteriumsvariable in die Berechnungen ein. Basis für die Analysen sind die Itemparameter aus der ESNaS-Pilotierungsstudie von 2009.

3.3.1 Variablen

Als Prädiktoren sollen das Vorhandensein von lösungsrelevanten Abbildungen in den Aufgabenstimuli und die schwierigkeitsgenerierenden Aufgabenmerkma-

le *Komplexität* und *kognitiver Prozess* in die Regressionsanalyse aufgenommen werden. Als Kriteriumsvariable dient die Aufgabenschwierigkeit in Form des Itemparameters aus dem Rasch-Modell.

Störvariablen sind praktisch alle weiteren schwierigkeitsrelevanten Aufgabenmerkmale, die nicht in die Analyse eingehen. "Für viele [dieser] Aufgabenmerkmale hat es aus theoretischen Gründen wenig Sinn, spezielle kognitive Kompetenzen anzunehmen" (Prenzel et al., 2002, S. 129). Eine Sonderstellung nimmt unter diesen sicher das Antwortformat (offen vs. geschlossen) ein. Über die Frage, inwieweit unterschiedliche Antwortformate unterschiedliche Fähigkeiten erfordern, herrscht kein Konsens (Hollingworth, Beard & Proctor, 2007). Ein Einfluss dieser Eigenschaft auf die Itemschwierigkeit wäre insofern schwer interpretierbar und könnte kaum als Beleg für oder gegen eine Konfundierung gewertet werden. Auch Prenzel et al. (2002) führen dieses Merkmal als Beispiel für jene Kategorien an, die nicht "zur Charakterisierung der Itemschwierigkeit taugen" (S. 130). Da das Antwortformat in der vorliegenden Arbeit keinen Beitrag zur Beantwortung der behandelten Fragestellung leistet, bleibt es in den vorgenommenen Berechnungen unberücksichtigt.

Neben dem Antwortformat haben noch weitere "unterschiedliche theoretisch angenommene Prozesse beim Lösen, aber auch ... technische Oberflächencharakteristika" einen Einfluss auf die Aufgabenschwierigkeit (Hartig, 2007, S. 88). Entsprechend den Vorgaben bei der Aufgabenkonstruktion kann allerdings angenommen werden, dass die meisten dieser Eigenschaften weitgehend gleichmäßig über den großen Itempool verteilt sind, sodass ihr Einfluss auf die Ergebnisse der Analyse unerheblich ist.

3.3.2 Stichprobe

Die Stichprobe besteht im Fall der Regressionsanalyse aus den Items des Kompetenzbereichs *Erkenntnisgewinnung* im Fach Biologie. Für diesen Kompetenzbereich wurden im Rahmen der Pilotierung 189 Items im Multimatrix-Design von 2 829 Personen bearbeitet. Die Lösungen gingen in ein eindimensionales Raschmodell ein, auf dessen Basis Itemparameter (Logits) berechnet wurden. Von 189 Items dienten 27 der Erfassung von Kompetenzen im Bereich der wissenschaftstheoretischen Reflexion. Diese wurden (aus den in Kapitel 2.2.2 genannten Gründen) aus den Berechnungen ausgeschlossen.

Von den verbleibenden 162 Items enthielten 61 Items lösungsrelevante Abbildungen. Die Verteilung der Itemmerkmale *Komplexität* und *kognitiver Prozess* ist getrennt nach dem Vorhandensein lösungsrelevanter Abbildungen in Tabelle 3.1 dargestellt. In der Bedingung ohne Abbildungen sind einige Zellen unbesetzt, was insbesondere bei nominalskalierten Variablen zu erhöhten Standardfehlern der Regressionskoeffizienten führen kann (Menard, 2001, S. 78). Diesem Problem wird durch eine Umkodierung der betreffenden Itemmerkmale entgegengewirkt (vgl. nächster Abschnitt). Des Weiteren ist zu prüfen, ob die Prädiktorvariablen untereinander korreliert sind (Multikollinearität). Hohe Kollinearität kann zu insignifikanten Schätzern und hohen Standardfehlern führen (Menard, 2001, S. 76). Aus diesem Grund findet im Vorfeld der Regressionsanalysen jeweils eine Kollinearitätsdiagnose statt.

Tabelle 3.1: Häufigkeiten der Itemmerkmale *Komplexität* und *kognitiver Prozess* getrennt nach dem Vorhandensein lösungsrelevanter Abbildungen.

		kognitiver Prozess				
Format	Komplexität	Rep.	Sel.	Org.	Int.	gesamt
	1 Fakt	4	5	-	-	9
Text	2 Fakten	1	4	1	-	6
ohne	1 Zusammenhang	11	14	15	13	53
Abbildungen	2 Zusammenhänge	1	1	7	10	19
	überg. Konzept	0	1	2	11	14
	gesamt	17	25	25	34	101
	1 Fakt	5	0	-	-	5
Text	2 Fakten	0	3	0	-	3
mit	1 Zusammenhang	5	5	5	15	30
Abbildungen	2 Zusammenhänge	3	2	4	9	18
	überg. Konzept	1	0	0	4	5
	gesamt	14	10	9	28	61

Anmerkungen. Die Zellen *ein Fakt Organisieren*, *ein Fakt Integrieren* und *zwei Fakten Integrieren* sind standardmäßig unbesetzt.

Umkodierung.

Auf Basis der theoretischen Annahmen über ihren Einfluss auf die Aufgabenschwierigkeit könnten Komplexität und kognitive Prozesse zumindest als ordinalskaliert betrachtet werden. Verschiedene Studien (Ropohl, 2010; Wellnitz & Mayer, 2011) zeigen aber, dass sich diese Ordinalskalierung in den Daten nicht durchgängig bestätigt. Im Fall der Biologie-Items ist die Annahme der Ordinalskalierung jeweils auf den unteren beiden Niveaustufen der Komplexität und des kognitiven Prozesses verletzt. Um die beiden Variablen dennoch als Prädiktoren in das Regressionsmodell aufnehmen zu können, bestehen mehrere Möglichkeiten.

Eine Möglichkeit ist das Zusammenfassen der betroffenen Niveaustufen. Wenn die Abstände zwischen den verbleibenden Kategorien hierdurch zumindest annähernd gleich groß werden und eine lineare Beziehung zwischen den Kategorien und der Aufgabenschwierigkeit erreicht wird, würde man den Variablen dann Intervallskalierung unterstellen—ähnlich, wie dies häufig im Fall von Schulnoten getan wird (vgl. Bortz & Döring, 2006, S. 74; Klammer, 2005, S. 86). Dieses Vorgehen ist nicht unumstritten (vgl. Rost, 2005, S. 185). Mit einem Zusammenfassen der Niveaustufen wird auch dem Problem der unbesetzten Zellen entgegengewirkt.

Zwei weitere Möglichkeiten, mit dem Skalenniveau der beiden Merkmale umzugehen, sind die Umkodierung in Dummyvariablen (Bortz, 2005, S. 484) und die künstliche Dichotomisierung (Bortz, 2005, S. 226). In den Studien von Ropohl (2010) und von Wellnitz und Mayer (2011) zeigte sich, dass zwischen den jeweils unteren Niveaus der Variablen *Komplexität* und *kognitiver Prozess* keine statistisch bedeutsamen Schwierigkeitsunterschiede bestehen. Werden sie in Dummyvariablen umkodiert, dann werden für betreffenden Niveaus keine signifikanten Einflüsse auf die Aufgabenschwierigkeit gefunden (Ropohl, 2010, S. 87). Aus diesem Grund erscheint eine Dichotomisierung sinnvoller. Hierzu werden die Kategorien derart zusammengefasst, dass sich jeweils nur noch zwei Ausprägungen ergeben.

Da die Behandlung ordinalskalierter Variablen als intervallskaliert einerseits und die künstliche Dichotomisierung andererseits jeweils mit unterschiedlichen messtheoretischen Problemen einhergehen, wurde entschieden, beide Ansätze zu verfolgen und dementsprechend zwei verschiedene Regressionsmodelle zu

rechnen. Für das erste Modell werden die jeweils unteren beiden Niveaustufen der Variablen *Komplexität* und *kognitiver Prozess* zusammengefasst und die Variablen in dieser Form als Prädiktoren in die Regressionsanalyse aufgenommen. Für das zweite Modell werden die Komplexität und die kognitiven Prozesse nach inhaltlichen Kriterien in jeweils zwei Gruppen aufgeteilt. Für die Variable *Komplexität* beinhaltet die erste Gruppe alle Items, in denen nur einzelne (verbundene oder nicht verbundene) Fakten verarbeitet werden müssen. Hierzu werden die Komplexitätsniveaus *ein Fakt*, *zwei Fakten* und *ein Zusammenhang* zusammengefasst. Die zweite Gruppe beinhaltet alle Items, deren Lösung die Verarbeitung komplexerer Informationen erfordert. Dies entspricht den Komplexitätsniveaus *zwei Zusammenhänge* und *übergeordnetes Konzept*. Bei den kognitiven Prozessen beinhaltet die erste Gruppe alle Items, zu deren Lösung praktisch nur Informationen aus dem Aufgabenstimulus entnommen werden müssen. Dies entspricht den Prozessen *Reproduzieren* und *Selegieren*. Die zweite Gruppe beinhaltet Items, zu deren Lösung komplexere kognitive Operationen ausgeführt werden müssen. Dies entspricht den Prozessen *Organisieren* und *Integrieren*.

3.4 Ergebnisse

Für das erste Regressionsmodell wurden jeweils die unteren beiden Niveaustufen der Variablen *Komplexität* und *kognitiver Prozess* zu einer Stufe zusammengefasst (kollabiert). Vor dem Rechnen der Regressionsanalyse wurde das Modell auf mögliche Multikollinearität überprüft. Zur Überprüfung wurden sowohl die Toleranzwerte als auch der Konditionsindex herangezogen.

Toleranzwerte kleiner als .20 werden als Hinweis auf Multikollinearität angesehen (Menard, 2001, S. 76). Mit Werten zwischen .75 und .99 liegen die empirischen Toleranzwerte des ersten Regressionsmodells deutlich außerhalb des als kritisch angesehenen Bereichs.

Anhand der Konditionsindizes kann die Stärke einer möglichen Multikollinearität eingeschätzt werden. Werte zwischen 10 und 30 gelten als moderat bis stark und Werte über 30 als sehr stark (Janssen & Laatz, 2007, S. 433). Die Überprüfung der Konditionsindizes ergibt für das erste Modell Werte zwischen 1.00 und 7.51, sodass davon ausgegangen werden kann, dass die Analysen nicht

wesentlich durch Multikollinearitätsprobleme beeinträchtigt sind.

Das erste Regressionsmodell klärt 20% der Gesamtvarianz auf. Es zeigen sich deutliche Effekte der Aufgabenmerkmale auf die Schwierigkeit (Tabelle 3.2). Diese steigt wie erwartet mit zunehmender Komplexität und mit dem zum Lösen erforderlichen kognitiven Prozess linear an. Die Vorgabe von lösungsrelevanten Abbildungen im Aufgabenstimulus führt zu einer Reduktion der Itemschwierigkeit.

Tabelle 3.2: Lineare Regressionsanalyse der Aufgabenmerkmale zur Vorhersage von Itemschwierigkeiten (Regressionsgewichte, Standardfehler und standardisierte Beta-Koeffizienten)

Aufgabenmerkmale	*B*	*SE*	β
Komplexität	.492	.156	$.260^{**}$
kognitiver Prozess	.419	.151	$.229^{**}$
lösungsrelevante Abbildungen	–.547	.240	$–.163^{*}$

Anmerkungen. Simultane Regressionsanalyse. $R^2 = .199$. $^{*}p < .05$. $^{**}p < .01$.

Werden die Variablen *Komplexität* und *kognitiver Prozess* künstlich dichotomisiert (zweites Modell), ergeben sich für diese beiden Merkmale infolge der gröberen Abstufung größere unstandardisierte Regressionsgewichte als im ersten Modell; die standardisierten Regressionsgewichte fallen wiederum ähnlich aus wie im ersten Modell. Der Einfluss des Vorhandenseins von lösungsrelevanten Abbildungen ist in beiden Modellen nahezu identisch. Das zweite Modell Modell klärt 18% der Gesamtvarianz auf (Tabelle 3.3).

3.5 Diskussion

Die Ergebnisse der Regressionsanalyse fallen hypothesenkonform aus. Die gefundenen Einflüsse der Variablen *Komplexität* und *kognitiver Prozess* entsprechen den Erwartungen, die aus dem Aufgabenmodell abgeleitet wurden (Kauertz et al., 2010). Dieser Befund lässt sich im Sinne einer konvergenten Validierung als Beleg für die Validität des Modells interpretieren.

Tabelle 3.3: Lineare Regressionsanalyse der Aufgabenmerkmale zur Vorhersage von Itemschwierigkeiten bei künstlicher Dichotomisierung der Merkmale *Komplexität* und *kognitiver Prozess* (Regressionsgewichte, Standardfehler und standardisierte Beta-Koeffizienten)

Aufgabenmerkmale	*B*	*SE*	β
Komplexität (dichotomisiert)	.735	.265	.215**
kognitiver Prozess (dichotomisiert)	.876	.257	.265**
lösungsrelevante Abbildungen	–.538	.243	–.160*

Anmerkungen. Simultane Regressionsanalyse. $R^2 = .178$. $^{*}p < .05$. $^{**}p < .01$.

Auch die Hypothese, dass sich Abbildungen positiv auf die Lösungswahrscheinlichkeit auswirken, konnte für die ESNaS-Aufgaben des Bereichs *Erkenntnisgewinnung* im Fach Biologie bestätigt werden. Im Sinne einer diskriminanten Validierung kann dieser Befund als Hinweis auf eine Konfundierung der Testleistungen mit dem Textverständnis der Versuchspersonen interpretiert werden (Prenzel et al., 2002).

Aufgrund des quasi-experimentellen Designs und der Tatsache, dass es vermutlich weitere schwierigkeitsrelevante Aufgabenmerkmale gibt, die in der Analyse unberücksichtigt blieben, lassen die Ergebnisse allerdings keine zuverlässigen Aussagen über die Prozesse zu, die dem gefundenen Schwierigkeitsunterschied zugrunde liegen. Der Multimedia-Effekt (Mayer, 1997) und das ITPC-Modell (Schnotz, 2005) liefern mögliche Erklärungsansätze. Um diese Ansätze zu prüfen, ist ein experimentelles Untersuchungsdesign nötig, in dem das Darbietungsformat systematisch manipuliert wird und die anderen Aufgabenmerkmale gleichzeitig konstant gehalten werden. Um Hinweise auf die Rolle von Leseleistungen zu erhalten, sollten bei der Analyse der Ergebnisse des Biologie-Tests zusätzlich mögliche Interaktionen aus dem Darbietungsformat der Aufgaben und dem Leseverständnis bzw. der Lesegeschwindigkeit der Versuchspersonen untersucht werden. Eine Studie, in der ein solcher experimenteller Ansatz verfolgt wird, ist im folgenden Kapitel beschrieben.

Kapitel 4

Experimentelle Studie

In Kapitel 2 wurden Hypothesen über mögliche Zusammenhänge zwischen dem Darbietungsformat, den Leseleistungen und den Ergebnissen von Tests zur Erfassung naturwissenschaftlicher Kompetenzen aufgestellt. In Kapitel 3 konnte ein Zusammenhang zwischen dem Vorhandensein von lösungsrelevanten Abbildungen und der Itemschwierigkeit in Aufgaben der ESNaS-Studie nachgewiesen werden. Die Ursachen für diesen Effekt und die Rolle bestimmter Personenmerkmale bei dessen Zustandekommen sollen in einer experimentellen Studie nun genauer untersucht werden.

Dazu werden Aufgaben aus dem ESNaS-Instrument entnommen und in zwei Versionen eingesetzt. In der einen Versuchsbedingung werden sämtliche Informationen in Form von Text präsentiert. In der anderen Versuchsbedingung werden alle unmittelbar lösungsrelevanten Bestandteile der Aufgabenstimuli in Form von Abbildungen dargestellt. Von diesem Unterschied abgesehen sind die Aufgaben in beiden Versuchsbedingungen identisch. Auf diese Weise kann zuverlässig eingeschätzt werden, inwieweit Schwierigkeitsunterschiede tatsächlich auf das Darbietungsformat der lösungsrelevanten Informationen zurückzuführen sind. Außerdem kann verglichen werden, ob sich Abbildungen eignen, um eine Konfundierung der Ergebnisse mit sprachlichen Fähigkeiten zu reduzieren.

Die Untersuchung der Forschungsfrage lässt sich auf Basis eines vergleichsweise kleinen Itemsatzes realisieren und kann mit Methoden der klassischen Testtheorie ausgewertet werden. Hierzu ist zunächst die Konstruktion eines Instrumentes nötig, das hinsichtlich der Schwierigkeit und Trennschärfe der Items sowie der internen Konsistenz der Skalen in beiden Versionen des Tests

zufriedenstellende Werte aufweist. Um dieses Ziel zu erreichen, waren vor der eigentlichen Untersuchung zwei Vorstudien erforderlich. Da diese Vorstudien und die Hauptuntersuchung nach einer gemeinsamen Methodik konzipiert wurden, ist diese zusammenfassend für alle drei Erhebungen in Kapitel 4.1 dargestellt. Im Rahmen der Optimierung des Instruments ergab sich allerdings auch eine Reihe von Unterschieden zwischen den einzelnen Studien. Diese Ergänzungen und Anpassungen sind in separaten Methodenteilen in den Kapiteln zur jeweiligen Untersuchung beschrieben.

4.1 Methode

4.1.1 Unabhängige und abhängige Variablen

Die unabhängige Variable in den Vorstudien und der Hauptuntersuchung ist das Darbietungsformat der lösungsrelevanten Informationen, das in den Testaufgaben zur Erfassung biologischer Kompetenz in Erkenntnisgewinnung zum Einsatz kommt. Die Variable kann zwei Ausprägungen annehmen: Text und statische Abbildungen. Da nur die lösungsrelevanten Informationen, nicht aber die anderen Aufgabenbestandteile in Abbildungen überführt werden, ergeben sich die beiden Versuchsbedingungen *Text ohne Abbildungen* und *Text mit Abbildungen*. Als abhängige Variable wird die Kompetenz im Bereich *Erkenntnisgewinnung* im Fach Biologie erfasst.

4.1.2 Moderatorvariablen

Es wird davon ausgegangen, dass eine Reihe von Personenmerkmalen einen Einfluss auf die untersuchten Effekte haben wird. Auf Basis der in Kapitel 2 vorgestellten Theorien und Befunde wurden das Leseverständnis, die Lesegeschwindigkeit, das individuelle Vor- und Weltwissen und eine mögliche Mehrsprachigkeit im Elternhaus der Versuchspersonen als potentielle Moderatorvariablen identifiziert sowie entsprechende Hypothesen über die zu erwartenden ATI-Effekte aufgestellt. Diese sollen im Rahmen einer Vorstudie untersucht und ggf. in die Hauptuntersuchung übernommen werden. Die Operationalisierung der Moderatorvariablen ist in den Methodenteilen der jeweiligen Studien beschrieben.

4.1.3 Störvariablen

Die Leistungen in schriftlichen Tests sind, abgesehen von den oben genannten Moderatorvariablen, vermutlich mit weiteren Eigenschaften und Fähigkeiten konfundiert, die im Rahmen der geplanten Untersuchung aus Gründen der Testökonomie nicht erfasst werden können und somit als Störvariablen berücksichtigt werden müssen. Zu den vermutlich einflussreichsten personenbezogenen Störvariablen zählen die kognitive Grundfähigkeit, der Cognitive Load und—im Fall von Aufgaben mit Abbildungen—die Fähigkeit des Bildverstehens. Als personenunabhängige Variable könnte zudem die Testzeit einen Einfluss auf das Zustandekommen von Gruppenunterschieden haben.

Kognitive Grundfähigkeit.

Die Leistungen in Testaufgaben zur Erfassung naturwissenschaftlicher Kompetenz dürften zu einem gewissen Maß mit kognitiver Grundfähigkeit konfundiert sein. "Die Korrelationen zwischen Intelligenz und Schulerfolg gehören zu den höchsten in der Psychologischen Diagnostik" (Prenzel et al., 2007, S. 132). Diese Zusammenhänge lassen sich zu einem nicht unerheblichen Teil auf die Rolle kognitiver Grundfähigkeiten beim Lernen zurückführen. So ist allgemein bekannt, "dass Intelligenz für domänenspezifische Wissenserwerbsprozesse – vor allem bei Anfängern – von großer Bedeutung ist" (Baumert et al, 2007, S. 124). Es ist jedoch auch anzunehmen, dass Intelligenz einen direkten Einfluss auf das Zustandekommen der Leistungen in der Testsituation hat: "Intelligentere Schülerinnen und Schüler können sich schneller auf neue Aufgaben einstellen, sie verfügen über effektivere Problemlösestrategien, erkennen leichter lösungsrelevante Regeln, verfügen über eine größere Verarbeitungskapazität und elaboriertere Gedächtnisstrategien" (ebd.). Es kann allerdings davon ausgegangen werden, dass die kognitive Grundfähigkeit eine deutlich geringere Rolle beim Zustandekommen der Leistungen im Naturwissenschafts-Test spielt als die Lesekompetenz (Leutner et al., 2004, S. 169).

Um einen Einfluss der kognitiven Grundfähigkeit auf die Ergebnisse der durchzuführenden Analysen weitestgehend auszuschließen, erfolgt die Zuordnung von Versuchspersonen zu den beiden Parallelformen des Tests in randomisierter Form. Dies wird durch abwechselndes Austeilen der beiden Testheft-

formen realisiert. Bei ausreichend großen Stichproben ist davon auszugehen, dass der Einfluss personenbezogener Störvariablen durch eine Randomisierung neutralisiert wird (Bortz & Döring, 2009, S. 54). Neben der kognitiven Grundfähigkeit wird durch die Randomisierung ein Einfluss weiterer Variablen, beispielsweise Motivation, ebenfalls weitestgehend ausgeschlossen.

Cognitive Load.

Da sich die beiden Parallelformen des Tests hinsichtlich des Darbietungsformates unterscheiden, muss davon ausgegangen werden, dass die Bearbeitung der Aufgaben unter Umständen eine unterschiedlich hohe kognitive Belastung für die Versuchspersonen darstellt. Unterschiedliche Darbietungsformate bzw. die Art, auf welche sie kombiniert werden, üben nachgewiesenermaßen einen Einfluss auf den Cognitive Load bei der Informationsverarbeitung aus (Chandler & Sweller, 1991). Für eine Kontrolle möglicher Effekte wäre es wünschenswert, diese Variable bei der Testung zusätzlich zu erfassen. Die anvisierte Testzeit von einer Schulstunde bietet jedoch kaum Spielräume für die Erfassung einer zusätzlichen Variable, da bereits naturwissenschaftliche Kompetenz und Leseverständnis / Lesegeschwindigkeit getestet werden müssen. Daher wurde entschieden, im Rahmen dieser Studie auf eine Erfassung des Cognitive Load zu verzichten.

Bildverstehen.

Die Testbedingung, in der die lösungsrelevanten Informationen als Abbildungen dargeboten werden, erfordert, dass die Probanden in der Lage sind, die eingesetzten Bilder zu verstehen. Daher ist nicht auszuschließen, dass die anvisierte Reduktion einer möglichen Konfundierung mit Leseverständnis mit einem Anstieg der Konfundierung mit Bildverstehen einhergeht. Um zu verhindern, dass die Informationsentnahme durch den Einsatz von Bildern erschwert wird, wird bei der Bildgestaltung auf Einfachheit und realitätsnahe Darstellung geachtet, sodass für die Verarbeitung der Informationen keine besonderen Kenntnisse, etwa bezüglich spezieller Darstellungskonventionen, erforderlich sind (s. Kapitel 4.1.4). Um sicher zu gehen, dass dies gelungen ist, soll die Verständlichkeit der Abbildungen im Rahmen der ersten Vorstudie erfasst werden.

Testzeit.

Die Testform *Text mit Abbildungen* erfordert geteilte Aufmerksamkeit, da visuell dargebotene Texte und Abbildungen nicht gleichzeitig verarbeitet werden können. Dies könnte sich in einer Erhöhung der Verarbeitungszeit niederschlagen (Mayer, 2005b; Schnotz, 2005). Im Fall der hier eingesetzten Aufgaben werden Abbildungen allerdings nicht als Ergänzung, sondern als Ersatz für (lösungsrelevante) Textabschnitte eingesetzt. Der Aufgabenstimulus kann in beiden Versuchsbedingungen größtenteils linear gelesen werden, sodass ein Hin- und Herspringen zwischen Text und Bildern weitgehend ausgeschlossen wird. Zudem sind die Aufgabentexte in der Bedingung *Text mit Abbildungen* kürzer als in den Aufgaben ohne Abbildungen, da ein Teil der lösungsrelevanten Informationen nicht im Text, sondern in den Abbildungen enthalten ist. Somit ist in der Bedingung *Text mit Abbildungen* weniger Text zu lesen, wodurch ggf. eine erhöhte Bearbeitungszeit ausgeglichen werden könnte.

Um sicherzustellen, dass für beide Parallelformen ausreichend Bearbeitungszeit zur Verfügung steht, soll in der ersten Vorstudie mit einer vergleichsweise geringen Itemzahl pro Testheft begonnen werden. Für die Original-Aufgaben des ESNaS-Projektes liegen Zeitschätzungen vor, sodass bereits beim Zusammenstellen der Testhefte tendenziell eingeschätzt werden kann, wie viele Items in der anvisierten Zeit bearbeitet werden können. Sollte sich die Testzeit als ausreichend erweisen, kann die Anzahl der verwendeten Items in den Folgestudien sukzessive gesteigert werden. Zusätzlich wird überprüft, ob sich die Parallelformen hinsichtlich der Anzahl nicht bearbeiteter Aufgaben unterscheiden.

4.1.4 Materialien

Erfassung biologischer Kompetenz im Bereich Erkenntnisgewinnung.

Die biologische Kompetenz im Bereich der Erkenntnisgewinnung wird mit Testitems erfasst, die im Rahmen des ESNaS-Projektes von mehreren Aufgabenentwicklerinnen und -entwicklern konstruiert worden sind. Die fachdidaktischen und messtheoretischen Modelle, die diesen Aufgaben zugrunde liegen, sind in Kapitel 2.2 beschrieben. Die lösungsrelevanten Informationen dieser Items wurden umkonstruiert, um jeweils zwei Formen (Text, statische Abbildungen) zu erhalten. Weitere Informationen im Aufgabenstimulus, die nicht unmittel-

bar lösungsrelevant sind, wurden in beiden Parallelformen des Tests in der Textform belassen.

Da die vorliegende Dissertation parallel zu den ersten Projektphasen von ESNaS angefertigt wurde, begann die Itemselektion und -umkonstruktion zu einem Zeitpunkt, zu dem noch keine Itemparameter zu den verwendeten Aufgaben vorlagen. Die Eignung einzelner Aufgaben für die geplante Studie konnte also bestenfalls anhand bestimmter Item-Eigenschaften geschätzt werden. Hierzu zählen der zu testende Kompetenzteilbereich, der zum Lösen erforderliche kognitive Prozess und die Komplexität der Aufgabe. Zu jedem Aufgabenstimulus und zu jedem Item existierte von Beginn an eine A-priori-Zeitschätzung des Entwicklers bzw. der Entwicklerin. Diese Eigenschaften wurden bei der Selektion berücksichtigt.

Parallel zur Anfertigung dieser Dissertation wuchs die Anzahl der fertig konstruierten ESNaS-Aufgaben an. Somit standen zu jedem Erhebungszeitpunkt neue Items für eine Auswahl und Umkonstruktion zur Verfügung. Im weiteren Verlauf des ESNaS-Projektes wurde zudem die in Kapitel 3 bereits erwähnte Pilotierungsstudie durchgeführt, sodass für einen Teil der ausgewählten Items ab einem gewissen Zeitpunkt empirische Daten wie Schwierigkeitsparameter und Trennschärfe vorlagen, die insbesondere bei der Konzeption der Hauptstudie Berücksichtigung fanden (siehe Kapitel 4.4).

Im Projekt ESNaS wurden für den Bereich *Erkenntnisgewinnung* im Fach Biologie 45 Aufgabenstämme mit insgesamt 189 Testitems konstruiert. Nicht jedes dieser Items eignete sich für den Einsatz in der hier vorliegenden Arbeit. Auf Basis von A-priori-Zeitschätzungen erwiesen sich einige Items hinsichtlich der Bearbeitungszeit von vorneherein als zu lang und somit als ungeeignet für den Einsatz in einem vergleichsweise kurzen Test.

Eine weitere wichtige Voraussetzung für die Auswahl war die Möglichkeit, die im Aufgabentext enthaltenen lösungsrelevanten Informationen in die Form einer Abbildung zu überführen. Hierbei sind gewisse Grenzen gesetzt. Schnotz (2005) verweist darauf, dass sich Gattungsbegriffe, abstrakte Konzepte und logische Operatoren in der Regel nicht eindeutig visualisieren lassen. Als Beispiel nennt er die Gattungsbegriffe “mammal”, “reptile” und den logischen Operator “or”. Diese lassen sich zwar problemlos in einem Satz unterbringen: “The Marsh Harrier feeds on mammals or reptiles”. Es ist jedoch unmöglich, diesen Satz in

eine einzige, eindeutige Abbildung zu überführen (S. 53).

Eine Sichtung der zu Beginn der Arbeit verfügbaren Items ergab, dass ein nicht unerheblicher Teil der Aufgabenstimuli derartige Begriffe oder Operatoren enthielt und somit nicht für eine Umkonstruktion infrage kam. Als weniger problematisch erwiesen sich hingegen Aufgaben, in denen die Versuchsaufbauten biologischer Untersuchungen beschrieben waren. Auch die Beschreibung von Kreisläufen und Modellen ließ häufig eine problemlose Überführung in Abbildungen zu. Wie viele Items in den Vorstudien und der Hauptuntersuchung zum Einsatz kamen, ist im Methodenteil der jeweiligen Studie beschrieben.

Ein weiteres Auswahlkriterium war das Antwortformat. Um auch bei einer vergleichsweise geringen Anzahl von Items eine ausreichende interne Konsistenz der Skala zu gewährleisten und eine zusätzliche Konfundierung der Ergebnisse mit der Artikulationsfähigkeit und -bereitschaft der Versuchspersonen auszuschließen, wurden ausschließlich geschlossene Items verwendet. Im Fall der ESNaS-Aufgaben handelt es sich dabei grundsätzlich um Items mit vier möglichen Antworten, von denen immer nur eine richtig ist.

Die Kompetenzen, die in den ESNaS-Teilbereichen *naturwissenschaftliche Untersuchungen* und *naturwissenschaftliche Modellbildung* beschrieben sind, weisen eine gewisse konzeptuelle Nähe zueinander auf (Mayer, 2007), sodass Aufgaben aus beiden Bereichen in die hier verwendeten Testhefte einflossen. Von der Selektion ausgeschlossen waren hingegen sämtliche Aufgaben zum Kompetenzteilbereich *wissenschaftstheoretische Reflexion*, da dessen Inhalte auf einer gänzlich anderen Ebene des Wissenschaftsverständnisses angesiedelt sind (vgl. Kap. 2.2).

Nachdem geeignete Items ausgewählt worden waren, erfolgte jeweils die Umkonstruktion in die beiden Formate *Text ohne Abbildungen* und *Text mit Abbildungen*. Sie erforderte mehrere Schritte. In einem ersten Schritt wurde der Aufgabenstimulus auf nicht benötigte Informationen untersucht. Im Projekt ESNaS folgen auf den Stimulus der meisten Aufgaben mehrere Testitems, die sich auf verschiedene Informationen in diesem Stimulus beziehen. In der hier vorliegenden Arbeit kamen jedoch in den meisten Fällen nur jeweils ein bis zwei Items pro Aufgabe zum Einsatz, sodass ein Teil der im Stimulus enthaltenen Informationen hinfällig war. Diese für das Lösen nicht relevanten Informationen wurden aus den Aufgabenstimuli entfernt. Selbiges gilt für schmückende

Textabschnitte und für bereits vorhandene Abbildungen, die keinen Bezug zu den Lösungen der zu bearbeitenden Testitems aufwiesen.

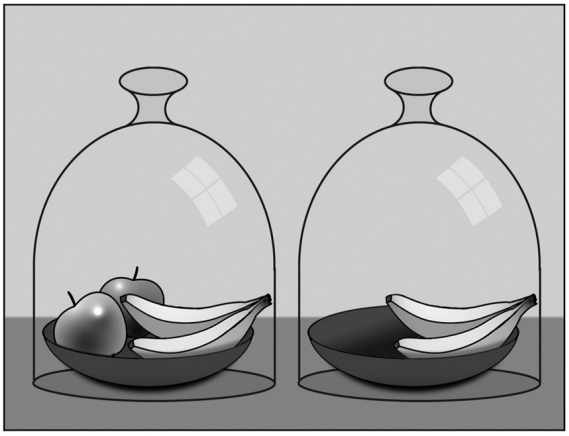
Im nächsten Schritt wurde analysiert, welche Information für das Lösen der Aufgabe relevant sind. Die hierzu notwendigen Angaben lassen sich größtenteils aus den Kodieranweisungen entnehmen, die im Rahmen der Aufgabenkonstruktion angefertigt wurden. Die lösungsrelevanten Informationen wurden für das Aufgabenformat *Text mit Abbildungen* vollständig in die Form einer oder mehrerer Grafiken überführt und der entsprechende Textabschnitt anschließend entfernt. Weitere Informationen im Aufgabenstimulus, die nicht unmittelbar lösungsrelevant sind, wurden in beiden Parallelformen des Tests in der Textform belassen.

Die Visualisierung erfolgte in Form realitätsnaher Schwarz-Weiß- bzw. Graustufen-Grafiken (vgl. Abbildung 4.1). Aufgrund ihrer starken Übereinstimmung mit Schemata der alltäglichen Wahrnehmung ist ein Erkennen der abgebildeten Gegenstände und Situationen möglich, ohne dass die Bilder spezielle Vorkenntnisse über Darstellungskonventionen von den Versuchspersonen fordern (Schnotz et al., 2010). Bei der Visualisierung wurde zudem darauf geachtet, dass beide Versionen informational äquivalent sind (Larkin & Simon, 1987; Peirce, 1906; Simon, 1978), d. h. in den Abbildungen sollten nicht mehr oder weniger Informationen enthalten sein als in der Text-Version. Außerdem wurde auf eine semantische Nähe (Kohärenz) sowie eine räumliche Nähe (Kontiguität) zusammengehöriger Informationen geachtet, um eine effiziente Verarbeitung zu gewährleisten (Mayer, 2003). Basierend auf diesen Kriterien entstanden von jeder Aufgabe zwei Versionen: Eine, die ausschließlich aus Text besteht, und eine weitere, in der die lösungsrelevanten Informationen nicht als Text, sondern als Abbildung dargestellt sind. Die Aufgaben der beiden Versuchsbedingungen unterscheiden sich also ausschließlich im Darbietungsformat der lösungsrelevanten Informationen und sind ansonsten identisch. Dementsprechend wurden pro Studie zwei Testhefte erstellt, die jeweils nur Aufgaben in einem der beiden Formate enthielten.

Obstkorb

Lisa dekoriert im Hause ihrer Eltern eine Obstschale, in welcher gelbe Bananen und reife Äpfel liegen. Nach zwei Tagen fällt ihr auf, dass die Bananen braun und fleckig geworden sind. Die Äpfel haben sich hingegen nicht verändert.

Um diese Beobachtung wissenschaftlich zu überprüfen, baut Lisa das folgende Experiment auf:



Anschließend beobachtet sie einige Tage lang, ob sich die Bananen in den beiden Obstschalen verfärben.

Welche Hypothese (Vermutung) will Lisa mit ihrem Experiment prüfen?

Obstkorb

Lisa dekoriert im Hause ihrer Eltern eine Obstschale, in welcher gelbe Bananen und reife Äpfel liegen. Nach zwei Tagen fällt ihr auf, dass die Bananen braun und fleckig geworden sind. Die Äpfel haben sich hingegen nicht verändert.

Um diese Beobachtung wissenschaftlich zu überprüfen, baut Lisa das folgende Experiment auf:

Sie legt 2 reife Äpfel und 2 Bananen in eine Obstschale und stülpt eine große Glasglocke darüber. In eine zweite Schale legt sie nur 2 Bananen, aber keine Äpfel. Auch über die zweite Obstschale wird eine Glasglocke gestülpt. Anschließend beobachtet sie einige Tage lang, ob sich die Bananen in den beiden Obstschalen verfärben.

Welche Hypothese (Vermutung) will Lisa mit ihrem Experiment prüfen?

Abbildung 4.1: Aufgabenstimulus und Fragestellung der Aufgabe "Obstkorb" in den beiden Versuchsbedingungen *Text mit Abbildungen* (oben) und *Text ohne Abbildungen* (unten). Abgesehen von den unterschiedlichen Darbietungsformaten der unmittelbar lösungsrelevanten Informationen sind die Texte identisch. Hinweis: Die Aufgabe ist nach dem ESNaS-Kompetenzmodell konstruiert, kam aber bei der Evaluation der Bildungsstandards nicht zum Einsatz.

Erfassung von Leseverständnis und Lesegeschwindigkeit.

Die psychologische Diagnostik und die kognitive Linguistik haben eine Reihe unterschiedlicher Verfahren hervorgebracht, um Leseleistungen zu messen (Hemminger et al., 2000). Viele davon sind als Einzeltests konzipiert, was ihre Eignung für den Einsatz in größeren Stichproben stark einschränkt (Schneider et al., 2007). Daneben existieren sowohl einige zur Gruppentestung geeignete Papier-und-Bleistift-Tests als auch computerbasierte Instrumente (Farcot & Latour, 2009).

Die Möglichkeit einer computerbasierten Testung wurde aufgrund der damit verbundenen infrastrukturellen Voraussetzungen für die hier vorgestellte Studie verworfen. Gängige, für eine Gruppentestung geeignete Papier-und-Bleistift-Verfahren sind Lückentexte (Schneider et al., 2007), Wortergänzungen (Klein-Braley, 1985) sowie Tests, in denen das Leseverständnis oder die Lesekompetenz der Versuchspersonen durch das Beantworten textinhaltsbezogener Fragen erfasst wird (Artelt et al., 2008).

Da die anvisierte Testzeit für die hier vorliegende Untersuchung bei einer Schulstunde lag, galt es, möglichst ökonomische Instrumente einzusetzen, die eine reliable und valide Messung in kurzer Zeit ermöglichen. Hierfür kamen in erster Linie Wortergänzungen oder Lückentexte infrage. Zusätzlich zum Leseverständnis sollte auch die Lesegeschwindigkeit erfasst werden.

Bei der Sichtung verfügbarer Instrumente zeigte sich, dass die meisten standardisierten Messinstrumente zur Erfassung von Leseverständnis auf die unteren Klassenstufen beschränkt sind, wohingegen kaum Tests vorhanden sind, die über das Grundschulalter hinausgehen (Hemminger et al., 2000). Eine Ausnahme bildet der LGVT 6-12 (Schneider et al., 2007), der mit einer Testzeit von insgesamt 10 Minuten eine reliable und ökonomische Testung, auch in größeren Gruppen, ermöglicht. Der LGVT 6-12 wurde im Rahmen der PISA-Erhebung des Jahres 2000 validiert und für die verschiedenen Schulformen normiert. Bei der Bearbeitung des Tests muss ein Text gelesen werden, der an mehreren Stellen die Auswahl eines geeigneten Wortes aus drei Alternativen erfordert. Auf diese Weise wird erfasst, wie präzise die Versuchsperson vorangegangene Textabschnitte gelesen hat. Zusätzlich wird die Lesegeschwindigkeit in Form der Anzahl von Wörtern erfasst, die in der festgelegten Testzeit von 4 Minuten gelesen wurden. Der Vergleich mit einer Reihe anderer Instrumente zeigt, dass

der LGVT 6-12 trotz seines sehr geringen Zeit- und Materialaufwandes gut geeignet ist, Leseverständnis valide zu erfassen (Schneider et al., 2007, S. 18). Mithilfe einer Ratekorrektur bei der Auswertung wird zudem Verfälschungstendenzen beim Bearbeiten erfolgreich entgegengewirkt (ebd.). Aufgrund seiner Eigenschaften hinsichtlich Messgüte und Ökonomie wird der LGVT 6-12 in allen hier vorgestellten Studien zur Erfassung von Leseverständnis und Lesegeschwindigkeit eingesetzt.

Personenmerkmale und Moderatorvariablen.

Zur Erfassung von Personenmerkmalen sowie der mutmaßlichen Moderatorvariablen Schulform, Schulnoten und Sprache im Elternhaus wurde ein einseitiger Personenfragebogen am Anfang der Biologie-Testhefte eingefügt. Die darin enthaltenen Variablen wurden von Studie zu Studie angepasst und sind in den jeweiligen Methodenteilen zu den einzelnen Untersuchungen beschrieben sowie im Anhang dokumentiert.

Anonymisierte Identifikation.

Um die erfasste Kompetenz in Erkenntnisgewinnung im Fach Biologie mit den Ergebnissen des LGVT 6-12 ins Verhältnis setzen zu können, bedurfte es einer eindeutigen Zuordnung der entsprechenden Testhefte zur jeweiligen Versuchsperson. Diese Zuordnung erfolgte über einen sechsstelligen Code, der sich aus den ersten beiden Buchstaben des Vornamens der Mutter, den ersten beiden Buchstaben des Vornamens des Vaters und dem Geburtstag der Testperson (Datum ohne Monat und Jahr) zusammensetzte. Die Schülerinnen und Schüler waren angewiesen, diesen Code auf beiden Testheften einzutragen.

4.1.5 Probanden

An den Vorstudien und der Hauptuntersuchung nahmen jeweils Schülerinnen und Schüler der Klassenstufe 9 teil. Es fanden Testungen in den Bundesländern Baden-Württemberg, Bremen, Hessen und Thüringen statt. Dabei wurden die Schultypen Hauptschule, Gesamtschule, Regelschule, Realschule und Gymnasium getestet. Die Zusammensetzung der einzelnen Stichproben ist in den Methodenteilen zu den jeweiligen Studien beschrieben.

4.1.6 Versuchsablauf

Die gesamte Untersuchung wurde jeweils innerhalb einer regulären Schulstunde (45 Minuten) durchgeführt. Die Bearbeitung der Testhefte zur Kompetenz im Bereich *Erkenntnisgewinnung* in Biologie erforderte 30 Minuten—inklusive der Instruktionen und eines Zeitpuffers für Nachfragen. Die Bearbeitung des LGVT 6-12 liegt inklusive Instruktionen und Zeit für Nachfragen bei 10 Minuten. Somit verblieben 5 Minuten für die Begrüßung und das Austeilen und Einsammeln der Testhefte. Die zeitliche Gestaltung erwies sich—von vereinzelten Ausnahmen abgesehen—für die überwiegende Mehrheit der Versuchspersonen als ausreichend, sodass der Test in allen Fällen vor dem Ertönen des Pausensignals beendet war. Eine Ausnahme bildet die Hauptstudie, in der sich die Testzeit für einen Teil der Schülerinnen und Schüler als zu kurz erwies. Dieser Umstand wurde bei der Auswertung der Daten berücksichtigt.

Zu Beginn der Schulstunde erfolgte die Begrüßung der Anwesenden durch die jeweilige Versuchsleiterin bzw. den jeweiligen Versuchsleiter sowie eine kurze Information über den Ablauf. Die Schülerinnen und Schüler wurden informiert, dass die Teilnahme freiwillig und anonym erfolgt, dass die Ergebnisse keinen Einfluss auf die Schulnoten haben und dass eine etwaige Verweigerung der Teilnahme keine negativen Auswirkungen hat. Daraufhin wurden die Testhefte für Erkenntnisgewinnung in Biologie ausgegeben. Durch ein abwechselndes Austeilen der beiden Testheftformen wurde eine zufällige Zuordnung der Personen zur jeweiligen Versuchsbedingung gewährleistet und gleichzeitig einem Abschreiben von den Sitznachbarn entgegengewirkt.

Nachdem jede Person ein Testheft erhalten hatte, wurden die Instruktionen auf dem Deckblatt von der Versuchsleiterin bzw. vom Versuchsleiter laut vorgelesen. Anschließend wurden die Schülerinnen und Schüler aufgefordert, die erste Seite aufzuschlagen und ihre Daten einzutragen. Die Zusammensetzung des persönlichen Identifizierungs-Codes wurde beispielhaft an der Tafel vorgeführt. Schülerinnen und Schüler, denen der Name eines oder beider Elternteile nicht bekannt war, wurden gebeten, stattdessen die Buchstaben XX einzutragen. In einem Fall waren Zwillinge unter den Versuchspersonen, deren Code identisch war. Ihre Testhefte wurden beim Einsammeln vom Versuchsleiter markiert, um bei der Dateneingabe eine eindeutige Zuordnung zu ermöglichen.

Nachdem alle Personenbögen vollständig ausgefüllt waren, wurden die Schü-

lerinnen und Schüler informiert, dass sie bei Problemen wie schlecht lesbaren Kopien die Aufsichtsperson um Hilfe bitten können, bei inhaltlichen Fragen jedoch nicht. Sie wurden angewiesen, zügig und konzentriert zu arbeiten und keine Aufgabe zu überspringen. Anschließend wurden sie aufgefordert, mit dem Bearbeiten der Aufgaben zu beginnen. Die Bearbeitung wurde nach exakt 25 Minuten durch das Signal "Stopp! Bitte schlagt die Testhefte zu" beendet, und die Hefte wurden eingesammelt.

Parallel zum Einsammeln der Biologie-Testhefte wurde pro Person jeweils ein Exemplar des LGVT 6-12 ausgeteilt. Die Schülerinnen und Schüler wurden aufgefordert, auf dem Bogen nur ihren persönlichen Identifizierungs-Code einzutragen und keine weiteren Eintragungen vorzunehmen. Anschließend wurde der LGVT 6-12 exakt nach den im Manual beschriebenen Instruktionen bearbeitet.

Nach Ablauf der Bearbeitungszeit wurden auch die Testhefte des LGVT 6-12 eingesammelt. Die Versuchsleiterin bzw. der Versuchsleiter dankte den Schülerinnen und Schülern für die Teilnahme und verabschiedete sich.

4.2 Erste Vorstudie

4.2.1 Ausgangssituation und Ziele der Studie

Die in diesem Kapitel dokumentierte Vorstudie diente vorrangig der Konstruktion eines zuverlässigen Instruments für die Hauptuntersuchung. Ein Vergleich der beiden Parallelformen setzt voraus, dass sie, abgesehen vom verwendeten Darbietungsformat, identisch sind, z. B. hinsichtlich der Anzahl und des Inhalts der verwendeten Items, der Aufgabenstimuli, der Fragestellungen und der Antwortmöglichkeiten. Diese Anforderungen, die es bei der Itemauswahl und -konstruktion zu beachten gilt, verursachen eine Reihe von Problemen.

Beispielsweise müssen beide Parallelformen des Instruments die für psychologische Tests geltenden Qualitätskriterien gleichermaßen zufriedenstellend erfüllen, insbesondere hinsichtlich der Schwierigkeit und Trennschärfe der verwendeten Items und der internen Konsistenz der Skala (Bortz & Döring, 2006, S. 218–222). Eine besondere Herausforderung ergibt sich dabei aus den prognostizierten Schwierigkeitsunterschieden zwischen den beiden Parallelformen.

Von Items, in denen die lösungsrelevanten Informationen als Abbildungen dargestellt sind, wird erwartet, dass sie leichter sind als Items, die ausschließlich in Textform dargeboten werden (vgl. Hypothese $H1_{Vorstudie1}$). Dabei soll der klassische Schwierigkeitsparameter aber keine Werte unter .20 bzw. über .80 erreichen. Sobald ein Item diese Anforderungen in einer der beiden Parallelformen nicht erfüllt, muss es aus beiden Versionen des Tests gestrichen werden. Es galt also, Items zu finden bzw. zu konstruieren, die sowohl in der Form *Text ohne Abbildungen* als auch in der Form *Text mit Abbildungen* eine Schwierigkeit zwischen .20 und .80 sowie eine zufriedenstellende Trennschärfe aufweisen.

Eine weitere Herausforderung stellte die informationale Äquivalenz der Parallelformen dar. Wenn Texte und Abbildungen informational äquivalent sind (Simon, 1978; Larkin & Simon, 1987; Schnotz, 2002), dann ist es nach dem Integrierten Modell des Text- und Bildverstehens von Schnotz (2005) möglich, auf Basis von Abbildungen zu denselben mentalen Repräsentationen zu gelangen, wie auf Basis von geschriebenen Texten. Informationale Äquivalenz stellt also eine notwendige Voraussetzung für einen zuverlässigen Vergleich der beiden Instrumente dar. Nur wenn diese Voraussetzung erfüllt ist, können gefundene Unterschiede zwischen den Parallelformen sicher auf das unterschiedliche Darbietungsformat (und nicht etwa auf Unterschiede im Informationsgehalt) zurückgeführt werden. Bei der Itemkonstruktion wurde deshalb streng darauf geachtet, dass in der Text- und in der Bildversion jeweils dieselben Informationen enthalten sind.

Informationale Äquivalenz kann aber nicht ausschließlich als eine Eigenschaft der Aufgaben gesehen werden. Ob ein Text und eine Abbildung dieselben Informationen enthalten, ist eine Frage, die sich streng genommen nur aus der Sicht der Rezipienten beantworten lässt. Aus diesem Grund muss sichergestellt werden, dass die Überführung der lösungsrelevanten Informationen aus den Aufgabentexten in Abbildungen auf eine zielgruppengerechte Weise erfolgt ist, d. h., dass die verwendeten Texte und Abbildungen für Schülerinnen und Schüler der 9. Klasse gleichermaßen zur Entnahme der relevanten Informationen geeignet sind.

4.2.2 Zu prüfende psychologische Hypothesen

Obwohl die erste Vorstudie primär der Konstruktion eines zuverlässigen Instruments und nur sekundär der Testung psychologischer Hypothesen diente, sollten auch erste Überprüfungen der zentralen Annahmen vorgenommen werden (unter Vorbehalt einer mindestens ausreichenden Güte von Items und Skala). Die im Rahmen dieser Vorstudie zu prüfenden Hypothesen wurden bereits in Kapitel 2.6 eingeführt und begründet:

$H1_{Vorstudie1}$: Mit kombinierten Text-Bild-Aufgaben werden bessere Leistungen erzielt als mit reinen Textaufgaben.

$H2_{Vorstudie1}$: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Leseverständnis und Darbietungsformat feststellbar.

$H3_{Vorstudie1}$: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Lesegeschwindigkeit und Darbietungsformat feststellbar.

Im Hinblick auf mögliche Interaktionen zwischen der erfassten Kompetenz in Biologie, dem Darbietungsformat der lösungsrelevanten Informationen und diversen Personenmerkmalen wird der Fokus in dieser Vorstudie zunächst auf die Sprache im Elternhaus und das individuelle Vor- und Weltwissen der Versuchspersonen gelegt:

$H4_{Vorstudie1}$: Schülerinnen und Schüler, in deren Elternhaus eine andere Sprache als Deutsch gesprochen wird, profitieren in höherem Maß vom Einsatz von Abbildungen als solche, die in deutschsprachigen Haushalten aufwachsen.

$H5_{Vorstudie1}$: Personen mit niedrigem Vorwissen profitieren in höherem Maß vom Einsatz von Abbildungen als Personen mit hohem Vorwissen.

Eine im Rahmen dieser Vorstudie zusätzlich eingeführte Hypothese nimmt Bezug auf die informationale Äquivalenz der beiden Parallelformen, bzw. deren Wahrnehmung durch die Rezipienten. Es wird postuliert, dass die verwendeten Texte und Abbildungen für die zu testenden Personen gleichermaßen zur Entnahme der lösungsrelevanten Informationen geeignet sind:

$H6_{Vorstudie1}$: Die Darbietungsformate werden von den Schülerinnen und Schülern als informational äquivalent wahrgenommen.

4.2.3 Methode

Variablen.

Die unabhängige Variable ist das Darbietungsformat der lösungsrelevanten Informationen in den Testaufgaben, mit denen biologische Kompetenz im Bereich der Erkenntnisgewinnung erfasst wird. Die Variable kann zwei Ausprägungen annehmen: Text und statische Abbildungen. Da nur die lösungsrelevanten Informationen, nicht aber die anderen Aufgabenbestandteile in Abbildungen überführt werden, ergeben sich die beiden Versuchsbedingungen *Text ohne Abbildungen* und *Text mit Abbildungen*. Als abhängige Variable dient die erzielte Punktzahl im Kompetenztest zur Erkenntnisgewinnung in Biologie. Moderierende Effekte werden vom Leseverständnis und von der Lesegeschwindigkeit, von einem ggf. vorhandenen fremdsprachigen Elternhaus und vom individuellen Vor- und Weltwissen der Schülerinnen und Schüler erwartet. Eine nur in dieser Vorstudie untersuchte abhängige Variable ist die Einschätzung der Schülerinnen und Schüler, in welchem Maß die Aufgabentexte bzw. die Abbildungen für das Lösen der Aufgabe hilfreich sind. Mögliche personenbezogene Störvariablen und die gewählten Maßnahmen zu ihrer Kontrolle sind in Kapitel 4.1 theoretisch begründet und diskutiert.

Materialien.

Zum Zeitpunkt der Itemselektion für diese Vorstudie lagen der fachdidaktischen Leitung des Faches Biologie 17 Aufgabenstämme mit insgesamt 89 Testitems zum Bereich *Erkenntnisgewinnung* aus dem ESNaS-Projekt vor. Es galt, geeignete Items für die Prüfung der im Rahmen dieser Arbeit aufgestellten

Hypothesen auszuwählen. Allgemeine Kriterien für die Selektion und Umkonstruktion sind in Kapitel 4.1 beschrieben.

Gemäß diesen Kriterien eigneten sich 14 der 89 Items für eine Umkonstruktion, von denen zudem sechs Items untereinander Unverträglichkeiten aufwiesen (d. h., die Inhalte eines Items gaben Hinweise für die Lösung eines anderen Items). Somit verblieben acht Items für den Einsatz in der Vorstudie.

In sechs der acht Aufgabenstimuli lagen die lösungsrelevanten Informationen in Form von Text vor. Die Überführung dieser Informationen in Abbildungen erfolgte in Form einfacher Schwarz-Weiß- bzw. Graustufengrafiken, die vom Autor selbst digital angefertigt wurden. Dabei wurde darauf geachtet, dass die Abbildungen möglichst schlicht gehalten waren und keine irrelevanten Informationen enthielten (siehe Beispielaufgabe in Abbildung 4.1). Es wurden nur die lösungsrelevanten Informationen in Abbildungen überführt. Für Aufgabenbestandteile ohne unmittelbare Lösungsrelevanz (z. B. Kontext und Fragestellung) wurde das Textformat beibehalten.

Zwei der acht ausgewählten Items enthielten bereits lösungsrelevante Informationen in Form von Abbildungen. Diese Items wurden in reine Textaufgaben umkonstruiert, sodass am Ende für jedes Item zwei Versionen vorlagen. Diesen beiden Versionen entsprechend wurden die beiden Parallelformen des Tests erstellt. Das Layout der Testhefte war so angelegt, dass jeweils nur ein Item pro Seite abgedruckt war. Zur Erfassung des Leseverständnisses und der Lesegeschwindigkeit der Versuchspersonen wurde der LGVT 6–12 eingesetzt (vgl. Kapitel 4.1.4).

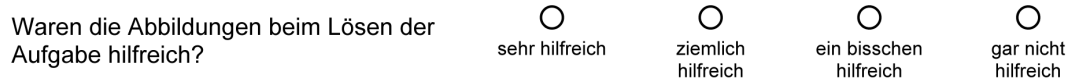


Abbildung 4.2: Auszug aus dem Testheft. Die abgebildete Frage schloss sich unmittelbar an jedes der zu bearbeitenden Items an (hier für die Versuchsbedingung *Text mit Abbildungen*). In der Versuchsbedingung *Text ohne Abbildungen* lautete die Formulierung "War der Aufgabentext beim Lösen der Aufgabe hilfreich?"

Um zu überprüfen, ob die Aufgabentexte und die Abbildungen den Schülerinnen und Schülern gleichermaßen zur Entnahme der lösungsrelevanten Informationen geeignet erscheinen, war jedem Item eine entsprechende Frage nachgeschaltet. Die Antwort erfolgte durch Ankreuzen auf einer vierstufigen Likert-Skala (Abbildung 4.2).

Neben dem Leseverständnis und der Lesegeschwindigkeit, die mit dem LGVT 6–12 erfasst worden sind, wurden moderierende Effekte eines ggf. vorhandenen fremdsprachigen Elternhauses sowie des individuellen Vor- und Weltwissens untersucht. Für Letzteres hat sich in zahlreichen Studien die Schulform als zuverlässiger Indikator erwiesen (Grube, Hartmann & Mayer, 2008; Köller, 2007; Trautwein, Köller, Lehmann & Lüdtke, 2007), weshalb sie hier zur Operationalisierung herangezogen wird. Die im Elternhaus gesprochenen Sprachen und die besuchte Schulform waren Bestandteil eines Personenfragebogens, der am Anfang des Kompetenztests in Biologie eingeheftet war (s. Anhang). Neben den bereits genannten Informationen wurden mit dem Fragebogen noch die folgenden personenbezogenen Variablen erfasst:

- Alter
- Geschlecht
- persönlicher Code zur Zuordnung der Testhefte

Räumlichkeiten.

Die Testung erfolgte in Klassenräumen der jeweils getesteten Schule. Die aus dem Unterricht bekannte Sitzordnung wurde—soweit ausreichend Platz zur Verfügung stand—so verändert, dass zwischen zwei Schülerinnen bzw. Schülern jeweils ein freier Platz lag.

Anwesende Personen.

Neben den Versuchspersonen war im Rahmen dieser Vorstudie nur die jeweilige Lehrkraft für das Fach Biologie anwesend. Sie übernahm auch die Testdurchführung.

Stichprobe.

An der ersten Vorstudie nahmen 115 Schülerinnen und Schüler der 9. Jahrgangsstufe aus Hessen teil (40 Hauptschülerinnen und Hauptschüler, 26 Realschülerinnen und Realschüler, 49 Gymnasiasten). Der Anteil weiblicher Versuchspersonen betrug 50%. Das Durchschnittsalter lag zum Testzeitpunkt bei 14.8 Jahren ($SD = 0.70$). Von den getesteten Personen gaben 30 an, dass in ihrem Elternhaus eine andere Sprache als Deutsch gesprochen wird; 11 Personen machten hierzu keine Angabe. Die häufigste Fremdsprache war Türkisch (13 Personen).

Ablauf.

Der Untersuchungsablauf erfolgte gemäß einer schriftlichen Instruktion, welche den Testleitern rechtzeitig vor der Durchführung zugesendet wurde (s. Anhang). Für eine detaillierte Beschreibung hinsichtlich Dauer, zeitlicher Abfolge, Instruktionen und Zuordnung der Testhefte sei auf die allgemeinen Ausführungen in Kapitel 4.1.6 verwiesen. Da die Versuchsleitung den Fachlehrerinnen und Fachlehrern übertragen wurde, waren diese aufgefordert, etwaige Auffälligkeiten bei der Testung zu notieren. Von allen Lehrkräften wurde darauf hingewiesen, dass die 25-minütige Testzeit für den Kompetenztest deutlich zu großzügig bemessen war. Mit wenigen Ausnahmen seien alle Schülerinnen und Schüler lange vor dem Ablauf der Testzeit mit der Bearbeitung fertig gewesen. Davon abgesehen gab es keine weiteren Hinweise.

4.2.4 Ergebnisse

Die Berechnungen zur ersten Vorstudie wurden mit SPSS 20 für Mac durchgeführt. Effektstärken wurden mit G*Power 3 für Mac (Faul et al., 2007) analysiert.

Test zur Erfassung biologischer Kompetenz in Erkenntnisgewinnung.

Die Kodierung der Antworten folgte den Richtlinien des ESNaS-Projekts. Dementsprechend wurde für richtig angekreuzte Antworten des Biologie-Tests jeweils ein Punkt vergeben; falsch angekreuzte und übersprungene Items wurden mit 0 Punkten gewertet. Die Skalenbildung erfolgte durch Aufsummieren

der richtigen Antworten. Die Testhefte ohne Abbildungen waren von 56 Personen bearbeitet worden; die Testhefte mit Abbildungen von 59 Personen. Beide Skalen wurden zunächst auf ihre interne Konsistenz geprüft. Im Fall von psychologischen Tests werden dabei häufig Werte von Cronbach's $\alpha \geq .70$ als akzeptabel angesehen (Field, 2009, S. 675); andere Quellen bezeichnen Werte $\geq .80$ als gut (vgl. Bortz & Döring, 2009, S. 199). Zur Berechnung von Gruppenvergleichen werden gelegentlich auch Werte zwischen .50 und .70 als ausreichend erachtet (Lienert & Raatz, 1998, S. 14). Unter Berücksichtigung der Tatsache, dass die hier verwendeten Items mehrere Aspekte eines vergleichsweise heterogenen Kompetenzkonstrukts abbilden, sollen Werte von $\geq .70$ als zufriedenstellend angenommen werden. Hinsichtlich der Trennschärfe werden Werte ab .30 als gut eingestuft (Field, 2009, S. 679).

Für die Bedingung *Text ohne Abbildungen* ergibt sich eine interne Konsistenz von Cronbach's $\alpha = .76$, und die Trennschärfe der Items kann, mit einer Ausnahme, als gut bezeichnet werden. Die interne Konsistenz der Skala *Text mit Abbildungen* ist mit $\alpha = .50$ hingegen nicht zufriedenstellend. Die Trennschärfe der Items bewegt sich in dieser Bedingung generell auf sehr niedrigem Niveau (sie liegt im Mittel nur bei .24). Die Itemschwierigkeiten liegen mit wenigen Ausnahmen im mittleren Bereich ($.20 \leq p_{\mathrm{i}} \leq .80$). In der Bedingung *Text ohne Abbildungen* erwies sich ein Item als zu leicht ($p_{\mathrm{i}} > .80$) und ein weiteres als zu schwer ($p_{\mathrm{i}} < .20$). In der Bedingung *Text mit Abbildungen* wies ein Item eine zu geringe Lösungswahrscheinlichkeit auf.

Auf Basis der Befunde zur Itemschwierigkeit und Trennschärfe wurden zwei Items aus den Skalen entfernt. Ein Item wurde trotz einer tendenziell zu geringen Schwierigkeit in der Bedingung *Text mit Abbildungen* in den Analysen beibehalten, da es in der Versuchsbedingung *Text ohne Abbildungen* eine mittlere Schwierigkeit und in beiden Versionen des Tests eine ausreichende Trennschärfe aufwies. Durch das Entfernen der beiden kritischsten Items konnte die unbefriedigende Reliabilität kaum verbessert werden; es ergeben sich nahezu gleichbleibende Werte (Cronbach's $\alpha = .75$ für die Skala *Text ohne Abbildungen* und $\alpha = .51$ für die Skala *Text mit Abbildungen*). Das Entfernen weiterer Items erwies sich als nicht zielführend, da eine leichte Besserung auf der einen Skala immer zu einer Verschlechterung auf der anderen führte. Die somit verbleibenden sechs Items wurden zur Bildung der Rohwertskalen aufsummiert.

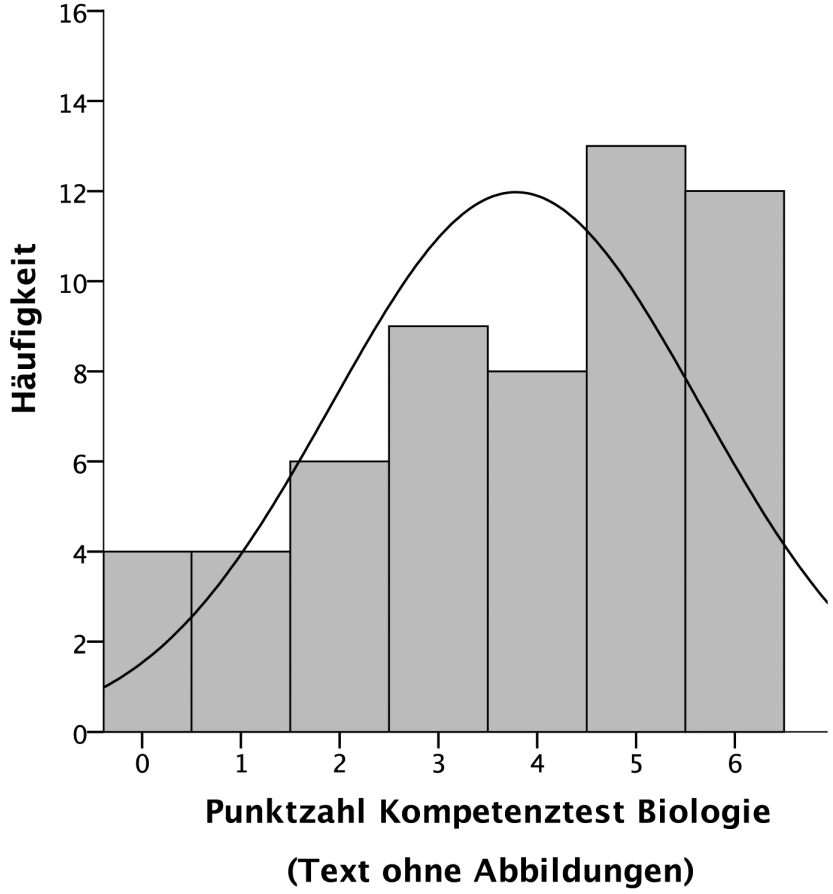


Abbildung 4.3: Histogramm für die Skala *naturwissenschaftiche Erkenntnisgewinnung in Biologie* in der Versuchsbedingung *Text ohne Abbildungen* (mit eingezeichneter Normalverteilungskurve). Es sind ein deutlicher Decken- und ein schwacher Bodeneffekt zu erkennen.

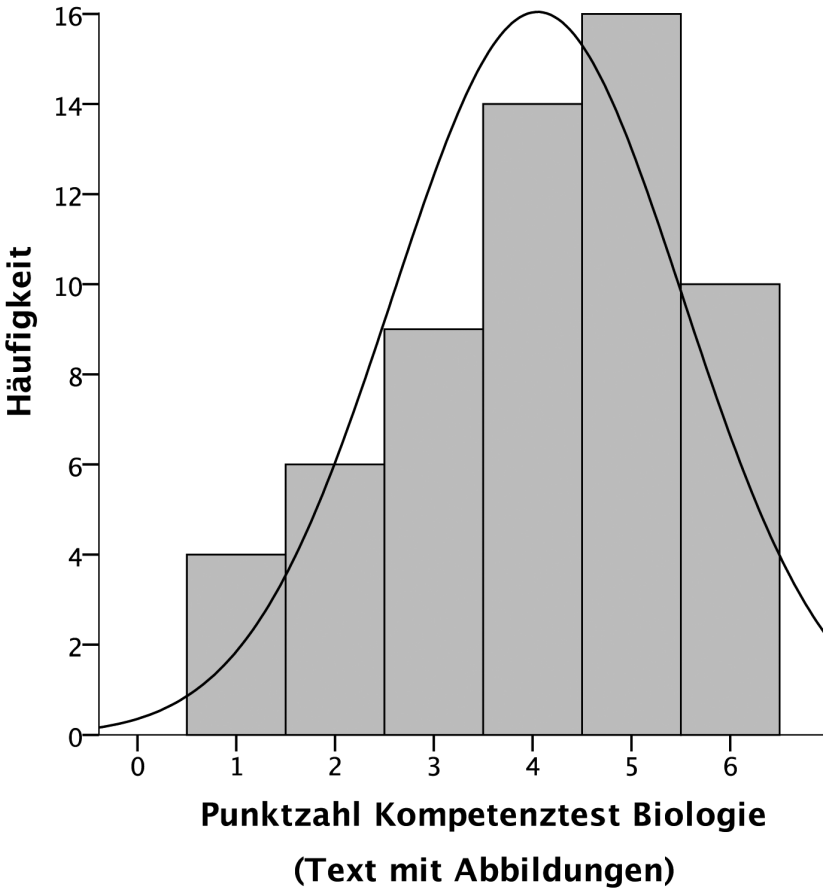


Abbildung 4.4: Histogramm für die Skala *naturwissenschaftiche Erkenntnisgewinnung in Biologie* in der Versuchsbedingung *Text mit Abbildungen* (mit eingezeichneter Normalverteilungskurve). Es ist ein deutlicher Deckeneffekt zu erkennen.

Die Skalen beider Versuchsbedingungen weisen erhebliche Deckeneffekte auf. In der Bedingung *Text ohne Abbildungen* zeigt sich zusätzlich ein schwacher Bodeneffekt. Beide Verteilungen weichen optisch deutlich von der Normalverteilung ab (Abbildungen 4.3 und 4.4). Die statistische Überprüfung (in Form der Tests nach Kolmogorow-Smirnow und Shapiro-Wilk) bestätigt jeweils, dass beide Verteilungen signifikant von der Normalverteilung abweichen ($p < .05$). Der Levene-Test ergibt zudem, dass keine Varianzhomogenität gegeben ist ($p < .05$).

LGVT 6–12.

Durch die randomisierte Zuordnung der Personen zu den beiden Versuchsbedingungen waren keine Gruppenunterschiede in den Leseleistungen zu erwarten. Um diese Annahme empirisch abzusichern, wurden die Ergebnisse des LGVT nach Gruppen getrennt ausgewertet und t–Tests für das Leseverständnis und die Lesegeschwindigkeit (mit dem Darbietungsformat des Biologie-Tests als Gruppenvariable) durchgeführt. Da hierbei nicht das Verwerfen, sondern die Bestätigung der Nullhypothese angestrebt ist, wird üblicherweise ein höheres Signifikanzniveau (z. B. .25 statt .05) verwendet (Bortz, 2005, S. 165).

Die ermittelten Werte entsprechen ungefähr den Normwerten, die im Anhang des LGVT-Manuals (Schneider et al., 2007) für die entsprechende Altersgruppe angegeben sind. Das arithmetische Mittel für das Leseverständnis liegt in der ersten Gruppe (Versuchsbedingung: *Text ohne Abbildungen*) bei 12.71 Rohwertpunkten ($SD = 9.19$) und in der zweiten Gruppe (Versuchsbedingung: *Text mit Abbildungen*) bei 12.46 Rohwertpunkten ($SD = 8.78$). Die mittlere Lesegeschwindigkeit in der vierminütigen Testzeit liegt in der ersten Gruppe bei 675.10 Wörtern ($SD = 226.22$) und in der zweiten Gruppe bei 692.18 Wörtern ($SD = 236.69$).

Für keine der beiden Variablen wurden signifikante Gruppenunterschiede gefunden; die Ergebnisse liegen mit $t_{\text{Leseverständnis}} = 0.15$, $p = .879$ (zweiseitig), $d = 0.03$ und $t_{\text{Lesegeschwindigkeit}} = -0.38$, $p = .705$ (zweiseitig), $d = 0.07$ jeweils deutlich über dem Signifikanzniveau von .25.

Multimedia-Effekt (Hypothese $H1_{Vorstudie1}$).

Die in den beiden Parallelformen erzielten Leistungen wurden trotz der Abweichung von der Normalverteilung mithilfe des t–Tests nach Student auf Mittelwertsunterschiede überprüft, da dieser in größeren Stichproben robust auf Verletzungen dieser Voraussetzung reagiert (Bortz, 2005, S. 141). Die deskriptiven Befunde suggerieren zunächst einen leichten Vorteil für die Versuchsbedingung mit Abbildungen im Vergleich zur Versuchsbedingung ohne Abbildungen. Der Effekt ist allerdings äußerst schwach und lässt sich in der getesteten Stichprobe nicht hinreichend gegen den Zufall absichern, sodass die Hypothese vorerst als nicht bestätigt gelten muss (Tabelle 4.1). Eine zusätzliche Prüfung mit einem nonparametrischen Verfahren (U–Test nach Mann & Whitney) ergab keine hiervon abweichenden Ergebnisse.

Tabelle 4.1: Vergleich der in den beiden Parallelformen erzielten Leistungen im Kompetenztest Biologie (deskriptive Werte, t–Test und Effektstärke)

Versuchsbedingung	*n*	*M*	*SD*	*df*	t_{emp}	p^*	*d*
ohne Abbildungen	56	3.79	1.87	104	–0.84	.200	0.16
mit Abbildungen	59	4.05	1.47				

*einseitiger Test.

Interaktionen aus Leseverständnis bzw. Lesegeschwindigkeit und Darbietungsformat (Hypothesen $H2_{Vorstudie1}$ und $H3_{Vorstudie1}$).

Gemäß den in Kapitel 2.4.2 beschriebenen theoretischen Annahmen wurde erwartet, dass der Effekt, den das Darbietungsformat auf die erzielten Leistungen im Biologie-Kompetenztest hat, mit den Leseleistungen der Versuchspersonen interagiert (ATI-Effekt). Die entsprechenden Hypothesen ($H2_{Vorstudie1}$ für die Interaktion aus Darbietungsformat und Leseverständnis sowie $H3_{Vorstudie1}$ für die Interaktion aus Darbietungsformat und Lesegeschwindigkeit) wurden mit ATI-Analysen überprüft, wobei die Werte der beiden LGVT-Skalen als Kovariaten in die varianzanalytischen Berechnungen eingingen. Bei der Interpretation der Ergebnisse ist zu berücksichtigen, dass die interne Konsistenz der Skala *Text mit Abbildungen* noch unbefriedigend ist und dass die Skala *Punktzahl*

im Kompetenztest Biologie in beiden Versuchsbedingungen von der Normalverteilung abweicht. Eine zufriedenstellende nonparametrische Alternative zu den durchgeführten Analysen besteht laut Field (2009) bisher nicht (S. 418).

Es wurde keine signifikante Interaktion aus Leseverständnis und Darbietungsformat festgestellt. Die Haupteffekte für Lesegeschwindigkeit und Darbietungsformat erwiesen sich ebenfalls als nicht signifikant. Ein signifikanter Haupteffekt ergab sich lediglich für das Leseverständnis (Tabelle 4.2).

Tabelle 4.2: Kovarianzanalyse für die Interaktion aus Leseverständnis und Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_{p}
Leseverständnis	55.78	1	55.78	28.51	$< .001$	.22
Lesegeschwindigkeit	0.06	1	0.06	0.03	.865	.00
Darbietungsformat	0.01	1	0.01	0.00	.950	.00
Leseverständnis $\times$ Darbietungsformat	0.45	1	0.45	0.23	.631	.00
Fehler	199.58	102	1.96	—	—	—

** Anmerkung.* Sequentielle Varianzzerlegung. $R^2 = .220$.

Tabelle 4.3: Kovarianzanalyse für die Interaktion aus Lesegeschwindigkeit und Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_{p}
Leseverständnis	55.78	1	55.78	28.51	$< .001$	.22
Lesegeschwindigkeit	0.06	1	0.06	0.03	.865	.00
Darbietungsformat	0.01	1	0.01	0.00	.950	.00
Lesegeschwindigkeit $\times$ Darbietungsformat	1.00	1	1.00	0.51	.476	.01
Fehler	199.03	102	1.95	—	—	—

** Anmerkung.* Sequentielle Varianzzerlegung. $R^2 = .222$.

Auch hinsichtlich der Interaktion aus Lesegeschwindigkeit und Darbietungsformat wurde kein signifikanter Effekt gefunden (Tabelle 4.3). Als statistisch signifikant erwies sich wiederum nur der Haupteffekt für das Leseverständnis; die Haupteffekte für die Lesegeschwindigkeit und das Darbietungsformat sind nicht signifikant. Die Hypothesen $H2_{Vorstudie1}$ und $H3_{Vorstudie1}$ müssen demnach vorerst als nicht bestätigt gelten.

Um einzuschätzen, wie stark die Ergebnisse des Biologie-Tests mit den Leseleistungen der Probanden zusammenhängen, wurden Korrelationen zwischen den erzielten Punktzahlen und den Ergebnissen des LGVT 6-12 berechnet. In beiden Versuchsbedingungen wurden mittlere Korrelationen zwischen der Kompetenz in Biologie und dem Leseverständnis gefunden. Der Korrelationskoeffizient beträgt .45 in der Bedingung ohne Abbildungen ($p < .001$) und .54 in der Bedingung mit Abbildungen ($p < .001$). Die Korrelationen mit Lesegeschwindigkeit fallen schwächer aus und erweisen sich als nicht signifikant von Null verschieden. Der Korrelationskoeffizient beträgt .14 in der Bedingung ohne Abbildungen ($p = .169$) und .02 in der Bedingung mit Abbildungen ($p = .444$). Die zugehörigen Streudiagramme sind in den Abbildungen 4.5 und 4.6 zu sehen.

Die gefundenen Werte wurden wegen der starken Deckeneffekte zusätzlich mit einem nonparametrischen Verfahren (Rangkorrelation nach Spearman) überprüft. Dabei ergaben sich keine wesentlich abweichenden Befunde.

Moderierende Effekte von Sprache und Vorwissen (Hypothesen $H4_{Vorstudie1}$ und $H5_{Vorstudie1}$).

Hinsichtlich möglicher Interaktionen zwischen der erfassten Kompetenz in Biologie, dem Darbietungsformat der lösungsrelevanten Informationen und diversen Personenmerkmalen wurde der Fokus in dieser Vorstudie vor dem Hintergrund der Testfairness zunächst auf die Sprache im Elternhaus und die Schulform (als Indikator für domänenspezifisches Vorwissen und allgemeines Weltwissen) gelegt. Die Überprüfung erfolgte trotz der nicht erfüllten Normalverteilungsannahme mittels zweifaktorieller ANOVA, da sich diese in größeren Stichproben als hinreichend robust erwiesen hat (Bortz, 2005, S. 286 f.).

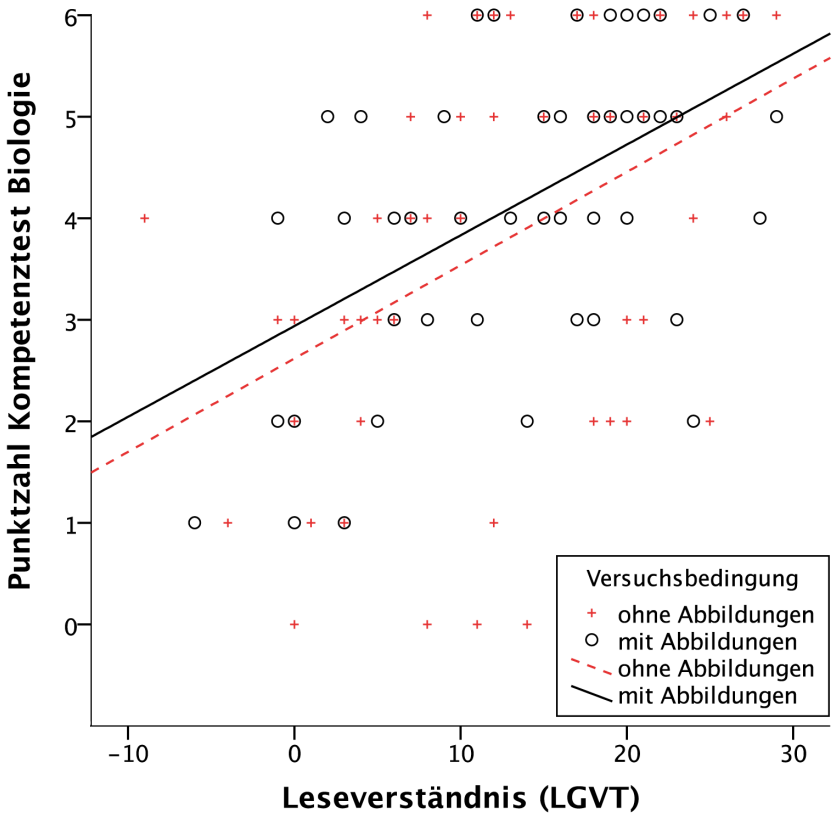


Abbildung 4.5: Streudiagramm mit Regressionsgeraden für den Zusammenhang zwischen der erzielten Punktzahl im Kompetenztest Biologie und dem Leseverständnis getrennt nach den beiden Versuchsbedingungen (*mit* und *ohne* Abbildungen). Eine Interaktion ist nicht festzustellen.

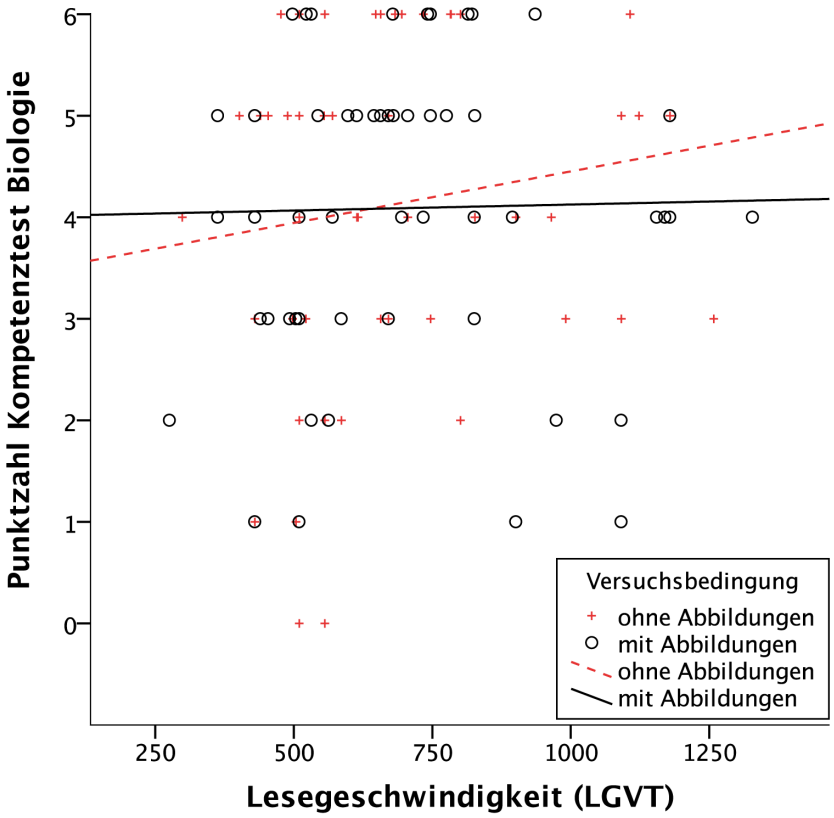


Abbildung 4.6: Streudiagramm mit Regressionsgeraden für den Zusammenhang zwischen der erzielten Punktzahl im Kompetenztest Biologie und der Lesegeschwindigkeit getrennt nach den beiden Versuchsbedingungen (*mit* und *ohne* Abbildungen). Es deutet sich eine schwache Interaktion an. Diese erweist sich jedoch als nicht signifikant.

Sprache im Elternhaus. Die Sprache, die im Elternhaus vorrangig gesprochen wird, hat einen Einfluss auf die Punktzahl, die im Biologie-Test erreicht wurde: Schülerinnen und Schüler aus deutschsprachigen Haushalten erzielten in der Versuchsbedingung ohne Abbildungen im Mittel 4.17 Punkte ($SD = 1.92$) und in der Versuchsbedingung mit Abbildungen 4.55 Punkte ($SD = 1.13$). Die Werte der Personen, in deren Elternhaus eine andere Sprache als Deutsch gesprochen wird, fallen niedriger aus: In der Versuchsbedingung ohne Abbildungen wurden im Mittel 3.00 Punkte erreicht ($SD = 1.41$), in der Versuchsbedingung mit Abbildungen lag der Mittelwert bei 3.29 Punkten ($SD = 1.64$).

Bereits die deskriptiven Werte lassen erkennen, dass fremd- oder mehrsprachig aufwachsende Personen von den eingesetzten Abbildungen nicht in höherem Maß profitieren konnten als Schülerinnen und Schüler, die in deutschsprachigen Haushalten leben (vgl. Abbildung 4.7). Die varianzanalytischen Befunde stützen dieses Ergebnis (Tabelle 4.4). Es wurde ein signifikanter Haupteffekt für die Sprache im Elternhaus (in den beiden Ausprägungen *deutsch* und *nicht deutsch*) gefunden. Daneben deutet sich ein schwacher Haupteffekt für das Darbietungsformat an, der sich allerdings als nicht signifikant erweist. Hinsichtlich einer Interaktion der beiden Variablen wurde kein signifikanter Effekt gefunden. Die Hypothese $H4_{Vorstudie1}$ wird durch die Daten demnach nicht bestätigt.

Tabelle 4.4: Varianzanalyse für die Interaktion aus der Sprache im Elternhaus und dem Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_p
Sprache im Elternhaus	32.38	1	32.38	13.42	$< .001$	.12
Darbietungsformat	3.31	1	3.31	1.37	.244	.01
Sprache im Elternhaus × Darbietungsformat	0.05	1	0.05	0.02	.882	.00
Fehler	241.25	100	2.41	—	—	—

Anmerkungen. *Sequentielle Varianzzerlegung. $R^2 = .129$.

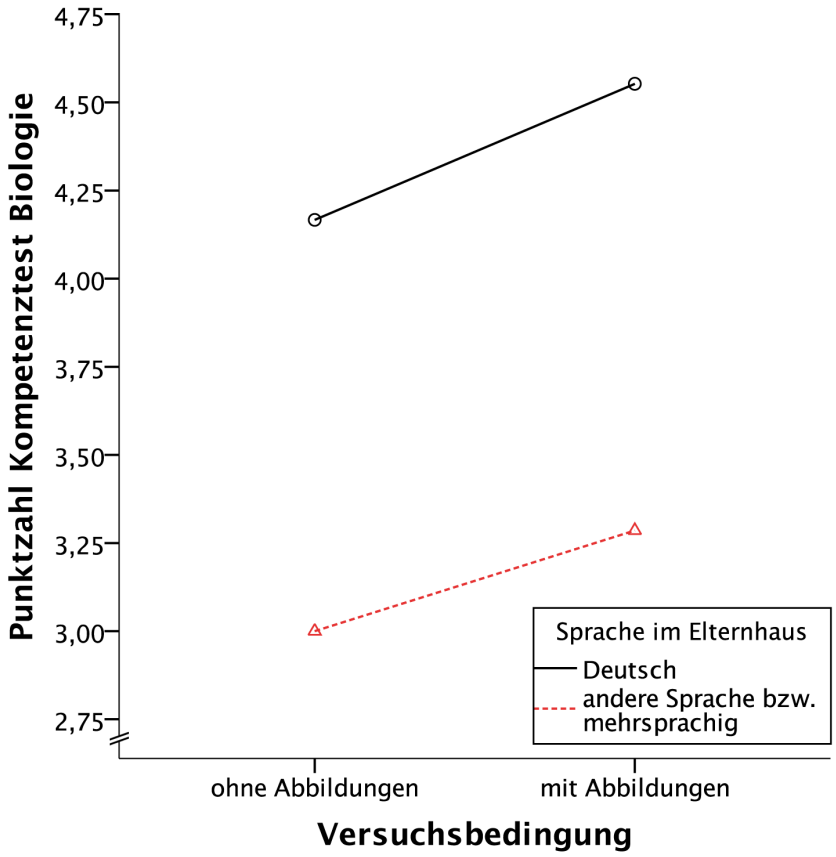


Abbildung 4.7: Der Haupteffekt der Sprache im Elternhaus erweist sich als statistisch bedeutsam. Eine Interaktion aus der Sprache im Elternhaus und dem Darbietungsformat wurde nicht gefunden.

Vorwissen. Hinsichtlich der Schulform, die hier als Indikator für das individuelle Vor- und Weltwissen herangezogen wurde, zeigt sich ein signifikanter Haupteffekt. Der Haupteffekt des Darbietungsformates ist nicht signifikant. Bei der Sichtung der deskriptiven Befunde deutet sich ein schwacher Interaktionseffekt an: Hauptschüler profitieren scheinbar am stärksten vom Einsatz von Abbildungen, wohingegen der Effekt bei Realschülern eher moderat ausfällt und sich bei den Gymnasiasten sogar leicht umzukehren scheint. Die Interaktion konnte allerdings nicht hinreichend gegen den Zufall abgesichert werden, sodass die Hypothese $H5_{\text{Vorstudie1}}$ vorerst als nicht bestätigt gelten muss. Eine zusammenfassende Darstellung der Analysen findet sich in Tabelle 4.5 und Abbildung 4.8.

Tabelle 4.5: Varianzanalyse für die Interaktion aus der Schulform und dem Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_{p}
Schulform	114.68	2	114.68	31.93	$< .001$	.37
Darbietungsformat	1.39	1	1.39	0.77	.382	.01
Schulform $\times$ Darbietungsformat	6.47	2	3.24	1.80	.170	.03
Fehler	195.76	109	1.80	—	—	—

Anmerkungen. *Sequentielle Varianzzerlegung. $R^2 = .385$.

Informationale Äquivalenz (Hypothese $H6_{\text{Vorstudie1}}$).

Es wurde postuliert, dass die verwendeten Texte und Abbildungen für die zu testenden Personen gleichermaßen zur Entnahme der lösungsrelevanten Informationen geeignet sind. Die Einschätzung der Lösungsrelevanz erfolgte anhand einer Einstufung auf einer vierstufigen Likert-Skala; die Prüfung der Hypothese wurde in Form eines t–Tests vorgenommen. Die Hypothese soll als bestätigt angesehen werden, wenn der gefundene Effekt entweder nicht existent (Null) oder vernachlässigbar klein ist ($d < 0.2$). Als Signifikanzniveau wird der von Bortz (2005, S. 165) vorgeschlagene Wert von .25 angesetzt.

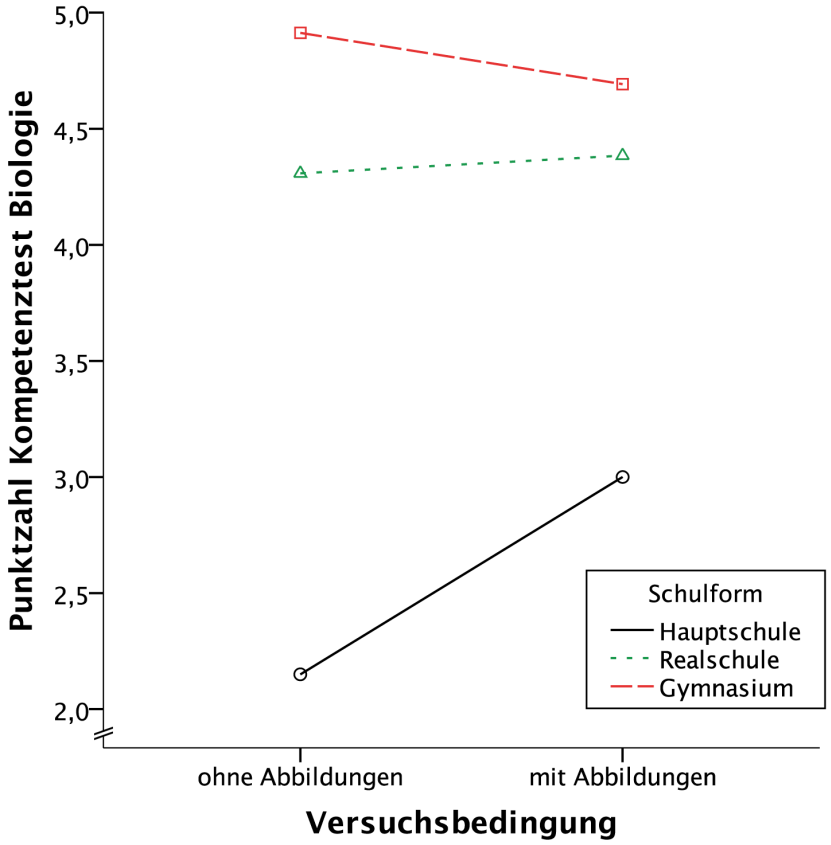


Abbildung 4.8: Hinsichtlich der Schulform (als Indikator für das Vor- und Weltwissen) wurde ein signifikanter Haupteffekt gefunden. Ein Interaktionseffekt aus Darbietungsformat und Schulform deutet sich optisch an, erweist sich aber als statistisch nicht bedeutsam.

Die Ergebnisse des Gruppenvergleichs zeigen, dass es keinen statistisch bedeutsamen Unterschied zwischen den beiden Versuchsbedingungen gibt (Tabelle 4.6). Die Aufgabentexte und die Abbildungen wurden von den Schülerinnen und Schülern als gleichermaßen hilfreich wahrgenommen. Die Hypothese $H6_{\text{Vorstudie1}}$ kann somit als bestätigt gelten.

Tabelle 4.6: Vergleich der wahrgenommenen Lösungsrelevanz des jeweiligen Darbietungsformates (deskriptive Werte, t–Test und Effektstärke)

Versuchsbedingung	n	M	SD	df	t_{emp}	p^*	d
ohne Abbildungen	52	2.22	0.67	105	0.70	.484	0.14
mit Abbildungen	55	2.31	0.66				

*zweiseitiger Test.

4.2.5 Diskussion

Das primäre Ziel der Vorstudie—die Konstruktion eines Instruments, das in beiden Versionen den Qualitätskriterien für psychologische Tests genügt—konnte nicht erreicht werden. Während die Versuchsbedingung *Text ohne Abbildungen* eine gute interne Konsistenz aufweist, erwies sich die Reliabilität der Skala *Text mit Abbildungen* als nicht zufriedenstellend. Da Items mit unbefriedigenden Kennwerten immer aus beiden Versionen des Tests entfernt werden müssen, ergab sich keine Möglichkeit, diese Skala ausreichend zu optimieren, ohne dass dabei gleichzeitig die Reliabilität der anderen Versuchsbedingung gesunken wäre.

Die verfügbare Bearbeitungszeit erwies sich als deutlich zu lang, sodass entschieden wurde, die Itemzahl bei zukünftigen Untersuchungen zu erhöhen. Mithilfe der Spearman-Brown-Formel (Moosbrugger, 2008b, S. 108) kann abgeschätzt werden, welchen Einfluss eine Testverlängerung auf die Reliabilität hätte. Eine gleichbleibende Trennschärfe der Items vorausgesetzt würde sich Cronbach's α durch ein Anheben der Itemzahl auf 12 in der Bedingung *Text mit Abbildungen* deutlich verbessern und würde dann bei .67 liegen. Durch den Einsatz von Items mit besserer Trennschärfe ließe sich der Wert möglicherweise auf ein zufriedenstellendes Niveau von .70 steigern. Eine noch weiter gehende

Verlängerung des Tests über die Anzahl von 12 Items hinaus erscheint vorerst nicht sinnvoll, da nicht sicher ist, ob die Bearbeitungszeit dafür ausreicht.

Als weiteres Problem erwies sich die Schwierigkeit des Kompetenztests zur Erkenntnisgewinnung in Biologie. In beiden Versuchsbedingungen wurden erhebliche Deckeneffekte gefunden, sodass die Skalen das vermutlich vorhandene Fähigkeitsspektrum im oberen Bereich nicht hinreichend abdeckten. Eine inhaltliche Interpretation der Ergebnisse der t–Tests, Korrelationen und Varianzanalysen ist angesichts der Probleme auf Skalen- und Itemebene nur sehr eingeschränkt möglich.

Die Ergebnisse zur Prüfung der Hypothesen $H1_{Vorstudie1}$, $H2_{Vorstudie1}$ und $H3_{Vorstudie1}$ sind aus den vorgenannten Gründen nur wenig belastbar. Bevor zuverlässige Aussagen über die Gültigkeit dieser Hypothesen möglich sind, bedarf es einer Optimierung des Instruments. Ähnliche Schlussfolgerungen sind auch für die Befunde zu den Moderatoreffekten von Sprache und Vorwissen zu ziehen. Auch hier sind die Ergebnisse infolge der geringen Reliabilität in der Bedingung *Text mit Abbildungen* nicht ausreichend belastbar. Von den messtheoretischen Problemen ausgenommen ist der Befund zu Hypothese $H6_{Vorstudie1}$. Die Daten zeigen, dass Schülerinnen und Schüler der 9. Klasse beide Darbietungsformate für das Lösen der Aufgaben als gleichermaßen hilfreich erachten. Dieses Ergebnis wird als Beleg dafür gewertet, dass es gelungen ist, beide Varianten des Tests informational äquivalent zu gestalten. Somit ist eine wichtige Voraussetzung für einen zuverlässigen Vergleich der beiden Parallelformen erfüllt. Die Erkenntnisse dieser ersten Vorstudie bilden die Basis für eine zweite Voruntersuchung, die in Kapitel 4.3 beschrieben ist.

4.3 Zweite Vorstudie

4.3.1 Ausgangssituation und Ziele der Studie

Bei der Auswertung der ersten Vorstudie ergaben sich messtheoretische Probleme infolge der unzureichenden Trennschärfe mehrerer Items zur Erfassung von Erkenntnisgewinnung in Biologie und der daraus resultierenden unbefriedigenden internen Konsistenz in der Bedingung *Text mit Abbildungen* sowie aufgrund der Deckeneffekte in beiden Versuchsbedingungen. Das primäre Ziel

der zweiten Vorstudie war aus diesem Grund die Fortführung der Bemühungen um die Konstruktion eines Instruments, welches die Qualitätsanforderungen für psychologische Tests in beiden Parallelformen zufriedenstellend erfüllt. Daneben sollten der in der ersten Vorstudie durchgeführte Leistungsvergleich zwischen den beiden Parallelformen und die ATI-Analysen zum Leseverständnis und zur Lesegeschwindigkeit wiederholt werden. Die Prüfung weiterer Moderatoreffekte (Vorwissen, Sprache im Elternhaus) wurde hingegen ausgesetzt, bis das Instrument ausreichend zuverlässig ist.

4.3.2 Zu prüfende psychologische Hypothesen

Wie bereits in der ersten Vorstudie, so stand auch in dieser Untersuchung die Konstruktion eines zuverlässigen Instruments im Vordergrund. Neben der obligatorischen Überprüfung von Skalen- und Itemkennwerten sollten folgende Hypothesen getestet werden:

$H1_{Vorstudie2}$: Mit kombinierten Text-Bild-Aufgaben werden bessere Leistungen erzielt als mit reinen Textaufgaben.

$H2_{Vorstudie2}$: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Leseverständnis und Darbietungsformat feststellbar.

$H3_{Vorstudie2}$: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Lesegeschwindigkeit und Darbietungsformat feststellbar.

4.3.3 Methode

Variablen.

Als unabhängige Variable wurde erneut das Darbietungsformat der lösungsrelevanten Informationen in den Testaufgaben zur Erfassung biologischer Kompetenz in Erkenntnisgewinnung manipuliert. Es liegt in den beiden Ausprägungen *Text* und *statische Abbildungen* vor. Als abhängige Variable dient wieder die erzielte Leistung im Kompetenztest zur Erkenntnisgewinnung in Biologie.

Zur Untersuchung mutmaßlicher Moderatoreffekte wurden das Leseverständnis und die Lesegeschwindigkeit erhoben. Die Erfassung des Vorwissens und der Sprache im Elternhaus wurde im Rahmen dieser Vorstudie hingegen zunächst ausgesetzt, da die Konstruktion eines zuverlässigen Instruments im Vordergrund stand. Mögliche Störvariablen wurden bereits in Kapitel 4.1 aufgeführt und diskutiert.

Materialien.

Unter Zuhilfenahme der Spearman-Brown-Formel wurde aus den Ergebnissen der ersten Vorstudie gefolgert, dass eine Verlängerung des Tests auf mindestens 12 Items nötig wäre, um eine befriedigende interne Konsistenz auf beiden Skalen zu erreichen. Zum Zeitpunkt der Itemselektion für diese Vorstudie lagen der fachdidaktischen Leitung des Faches Biologie 39 Aufgabenstämme mit insgesamt 179 Testitems zum Bereich *Erkenntnisgewinnung* aus dem ESNaS-Projekt vor. Aus diesem Itempool wurde erneut eine Auswahl nach den in Kapitel 4.1 beschriebenen Kriterien getroffen. Der überwiegende Teil der Items war bis dato noch nicht an ausreichend großen Stichproben überprüft worden, sodass keine verlässlichen Itemparameter für die Selektion herangezogen werden konnten.

Der Umfang der Testhefte wurde auf 12 Items erhöht. Unter diesen befanden sich auch sieben Items, die bereits in der ersten Vorstudie zum Einsatz gekommen waren und hinsichtlich Trennschärfe und Schwierigkeit zumindest in einer der beiden Versuchsbedingungen gute Parameter aufgewiesen hatten. Wie bereits in der ersten Vorstudie erfolgte wieder eine Umkonstruktion, sodass die lösungsrelevanten Informationen jeweils in den beiden Formaten *Text* und *statische Abbildungen* vorlagen. Zur Erfassung von Leseverständnis und Lesegeschwindigkeit kam erneut der LGVT 6–12 zum Einsatz.

Räumlichkeiten und anwesende Personen.

Die Testung erfolgte in den Klassenräumen der jeweils getesteten Schule. Die aus dem Unterricht bekannte Sitzordnung wurde—soweit ausreichend Platz zur Verfügung stand—so verändert, dass zwischen zwei Schülerinnen bzw. Schülern jeweils ein freier Platz lag. Neben den Versuchspersonen war im Rahmen

dieser Vorstudie nur die jeweilige Lehrkraft für das Fach Biologie anwesend. Sie übernahm auch die Testdurchführung.

Stichprobe.

An der zweiten Vorstudie nahmen 150 Schülerinnen und Schüler aus den Bundesländern Baden-Württemberg und Bremen teil (59 Realschülerinnen bzw. Realschüler und 91 Gymnasiasten). Der Anteil weiblicher Versuchspersonen betrug 47%. Das Durchschnittsalter lag zum Testzeitpunkt bei 15.0 Jahren ($SD = 0.66$). Von den getesteten Personen gaben 44 an, dass in ihrem Elternhaus eine andere Sprache als Deutsch gesprochen wird; eine Person machte hierzu keine Angabe. Häufigste Fremdsprache war erneut Türkisch (14 Personen).

Ablauf.

Der Untersuchungsablauf erfolgte gemäß einer schriftlichen Instruktion, welche den Testleiterinnen und Testleitern rechtzeitig vor der Durchführung zugesendet wurde (s. Anhang). Der genaue Ablauf (Dauer, zeitliche Abfolge, Instruktionen und randomisierte Zuordnung der Testhefte) ist in Kapitel 4.1.6 beschrieben.

Die Versuchsleiterinnen und Versuchsleiter wiesen erneut darauf hin, dass die anvisierte Testzeit für den Kompetenztest mit 25 Minuten zu großzügig bemessen war. Die Mehrheit der Schülerinnen und Schüler sei lange vor dem Ablauf der Testzeit mit der Bearbeitung fertig gewesen. Davon abgesehen gab es keine weiteren Hinweise.

4.3.4 Ergebnisse

Die Berechnungen zur zweiten Vorstudie wurden mit den Programmen SPSS 20 für Mac und G*Power 3 für Mac (letzteres zur Bestimmung von Effektstärken) durchgeführt.

Test zur Erfassung biologischer Kompetenz in Erkenntnisgewinnung.

Die Kodierung richtiger und falscher Antworten erfolgte nach den Richtlinien des ESNaS-Projekts. Für richtig angekreuzte Antworten des Biologie-Tests

wurde jeweils ein Punkt vergeben; falsch angekreuzte und übersprungene Items wurden mit 0 Punkten gewertet. Die Skalenbildung erfolgte durch Aufsummieren der richtigen Antworten.

Für die Bedingung *Text ohne Abbildungen* liegen 76 gültige Fälle vor. Die Skala weist eine interne Konsistenz von Cronbach's $\alpha = .63$ sowie einen leichten Deckeneffekt auf. Die rechnerische Überprüfung der Verteilung ergibt sowohl im Fall des K-S-Tests als auch mit dem strengeren Verfahren nach Shapiro-Wilk eine signifikante Abweichung von der Normalverteilung ($p < .001$).

Die Stichprobe für die Bedingung *Text mit Abbildungen* setzt sich aus 74 gültigen Fällen zusammen. Die interne Konsistenz dieser Skala beträgt $\alpha = .72$. Die Verteilung weicht optisch deutlich von der Normalverteilung ab. Auch diese Einschätzung wird rechnerisch sowohl durch den K-S-Test als auch durch die Überprüfung nach Shapiro-Wilk bestätigt ($p < .001$).

Durch das Entfernen der vier kritischsten Items aus beiden Bedingungen lässt sich die interne Konsistenz auf $\alpha = .66$ (Skala ohne Abbildungen) bzw. $\alpha = .75$ (Skala mit Abbildungen) erhöhen. Diese Maßnahme geht allerdings mit einer erheblichen Verstärkung der Deckeneffekte einher (Abbildungen 4.9 und 4.10). Im Test ohne Abbildungen weisen vier Items eine Lösungswahrscheinlichkeit von $p_i > .80$ auf; in der Bedingung mit Abbildungen ist hiervon ein Item betroffen. Da die betreffenden Items in beiden Bedingungen zufriedenstellende Trennschärfen aufwiesen, wurden sie jedoch zur Berechnung der geplanten Analysen beibehalten.

LGVT 6-12.

Das arithmetische Mittel für das Leseverständnis liegt bei 13.72 Rohwertpunkten ($SD = 6.97$) und die mittlere Lesegeschwindigkeit bei 774.18 Wörtern ($SD = 275.93$). Der Wert für das Leseverständnis liegt damit etwas unter dem Normwert für die getestete Altersgruppe; der Wert für die Lesegeschwindigkeit hingegen etwas über dem Wert, der im Manual publiziert ist. Da die Personen den beiden Versuchsbedingungen randomisiert zugeordnet wurden, waren keine Gruppenunterschiede in den Leseleistungen zu erwarten. Um diese Annahme abzusichern, wurden die Ergebnisse für das Leseverständnis und die Lesegeschwindigkeit nach Gruppen getrennt ausgewertet und t–Tests (mit dem Darbietungsformat des Biologie-Tests als Gruppenvariable) durchgeführt.

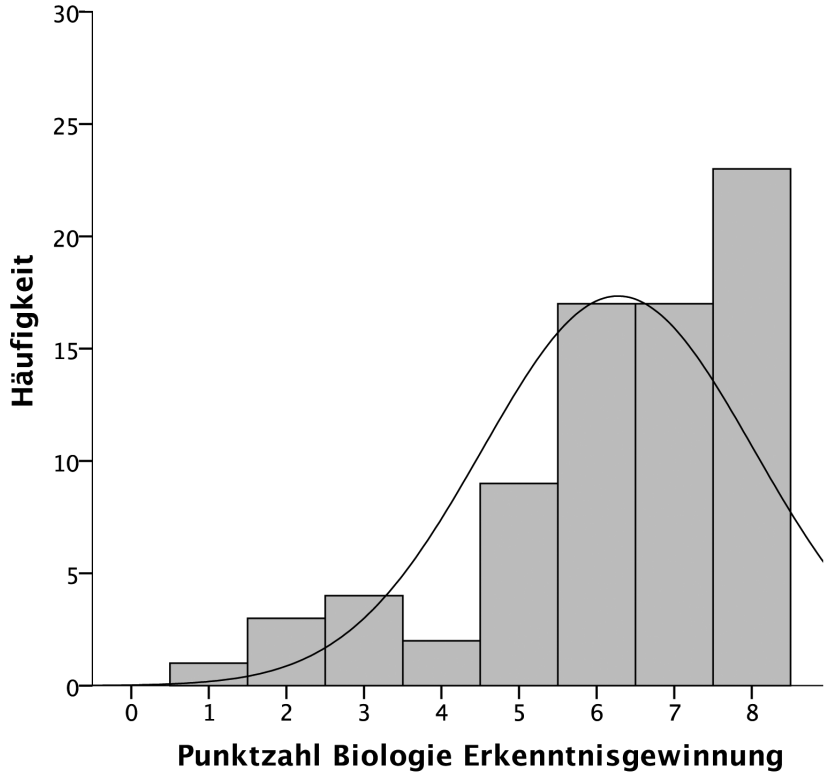


Abbildung 4.9: Histogramm für die Skala *naturwissenschaftiche Erkenntnisgewinnung in Biologie* in der Versuchsbedingung *Text ohne Abbildungen* (mit eingezeichneter Normalverteilungskurve). Es ist ein deutlicher Deckeneffekt zu erkennen.

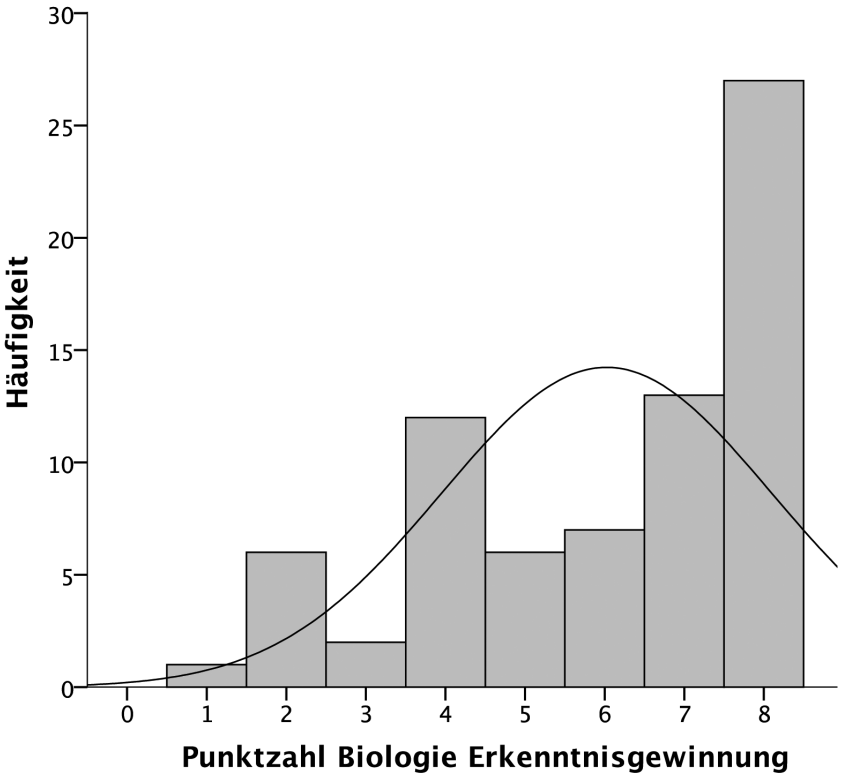


Abbildung 4.10: Histogramm für die Skala *naturwissenschaftiche Erkenntnisgewinnung in Biologie* in der Versuchsbedingung *Text mit Abbildungen* (mit eingezeichneter Normalverteilungskurve). Es ist ein deutlicher Deckeneffekt zu erkennen.

In der Versuchsbedingung *Text ohne Abbildungen* lag das Leseverständnis im Mittel bei 13.82 Rohwertpunkten (SD = 6.51); in der Bedingung *Text mit Abbildungen* wurden 13.62 Punkte erreicht (SD = 7.45). Es wurde kein signifikanter Unterschied gefunden; das Ergebnis liegt deutlich über dem von Bortz (2005, S. 165) für das Verwerfen von Nullhypothesen vorgeschlagenen Signifikanzniveau von .25 ($t_{\text{Leseverständnis}} = 0.17$, $p = .865$ [zweiseitig], $d = 0.03$).

Bei der Lesegeschwindigkeit deutet sich hingegen trotz Randomisierung ein schwacher Unterschied an. Die Schülerinnen und Schüler der Versuchsbedingung *Text ohne Abbildungen* wiesen mit durchschnittlich 737.71 gelesenen Wörtern (SD = 234.90) eine geringere Lesegeschwindigkeit auf als die Schülerinnen und Schüler der Bedingung *Text mit Abbildungen* (M = 811.64; SD = 390.69). Das Ergebnis des entsprechenden t–Tests stützt diesen Befund mit $t_{\text{Lesegeschwindigkeit}}$ = -1.64, p = .102 (zweiseitig). Die Effektstärke bewegt sich mit d = 0.27 im niedrigen Bereich. Hierzu sei allerdings angemerkt, dass ein nicht-signifikantes Ergebnis alleine kein Beleg für die Gültigkeit der Nullhypothese ist (Bortz & Döring, 2009, S. 650). Zudem sind nicht-signifikante Ergebnisse bei einem derart hohen α-Fehler-Niveau in größeren Stichproben relativ unwahrscheinlich, weshalb auch nicht alle Autoren das von Bortz vorgeschlagene Signifikanzniveau von .25 teilen, sondern stattdessen niedrigere kritische Werte vorschlagen (vgl. Bortz & Döring, 2009, S. 651).

Multimedia-Effekt (Hypothese $\mathbf{H1_{Vorstudie2}}$).

Es wurde prognostiziert, dass Aufgaben mit Abbildungen leichter zu lösen sind als Aufgaben ohne Abbildungen. Die Hypothese wurde trotz der Abweichung von der Normalverteilung mithilfe des t–Tests getestet, da dieser in größeren Stichproben robust auf Verletzungen dieser Voraussetzung reagiert (Bortz, 2005, S. 141).

In der getesteten Stichprobe wurde kein Effekt im Sinne der aufgestellten Hypothese gefunden (Tabelle 4.7). Das Ergebnis wurde zusätzlich mit einem nonparametrischen Verfahren (U–Test nach Mann & Whitney) überprüft, wobei sich keine abweichenden Befunde ergaben.

Tabelle 4.7: Vergleich der in den beiden Testformen erzielten Leistungen im Kompetenztest Biologie (deskriptive Werte, t-Test und Effektstärke)

Versuchsbedingung	*n*	*M*	*SD*	*df*	t_{emp}	p^*	*d*
ohne Abbildungen	76	6.28	1.75	143	0.80	.214	0.13
mit Abbildungen	74	6.03	2.07				

*einseitiger Test.

Interaktionen aus Leseverständnis bzw. Lesegeschwindigkeit und Darbietungsformat (Hypothesen $\mathbf{H2_{Vorstudie2}}$ und $\mathbf{H3_{Vorstudie2}}$).

Die prognostizierten Interaktionen aus dem Leseverständnis bzw. der Lesegeschwindigkeit und dem Darbietungsformat wurden erneut mit ATI-Analysen überprüft, wobei die Werte der beiden LGVT-Skalen als Kovariaten in die varianzanalytischen Berechnungen eingingen.

Eine signifikante Interaktion aus Leseverständnis und Darbietungsformat wurde nicht gefunden (Tabelle 4.8 und Abbildung 4.11). Es ergaben sich lediglich signifikante Haupteffekte für das Leseverständnis und die Lesegeschwindigkeit. Die Varianzaufklärung beträgt 22 Prozent.

Tabelle 4.8: Kovarianzanalyse für die Interaktion aus Leseverständnis und Darbietungsformat

Varianzquelle	SS^*	*df*	*MS*	F_{emp}	*p*	η^2_{p}
Leseverständnis	96.05	1	96.05	32.72	$< .001$	.18
Lesegeschwindigkeit	23.48	1	23.48	8.00	$< .01$	.05
Darbietungsformat	0.32	1	0.32	0.11	.742	.00
Leseverständnis × Darbietungsformat	0.00	1	0.00	0.00	.982	.00
Fehler	425.63	145	2.94	—	—	—

**Anmerkung.* Sequentielle Varianzzerlegung. $R^2 = .220$.

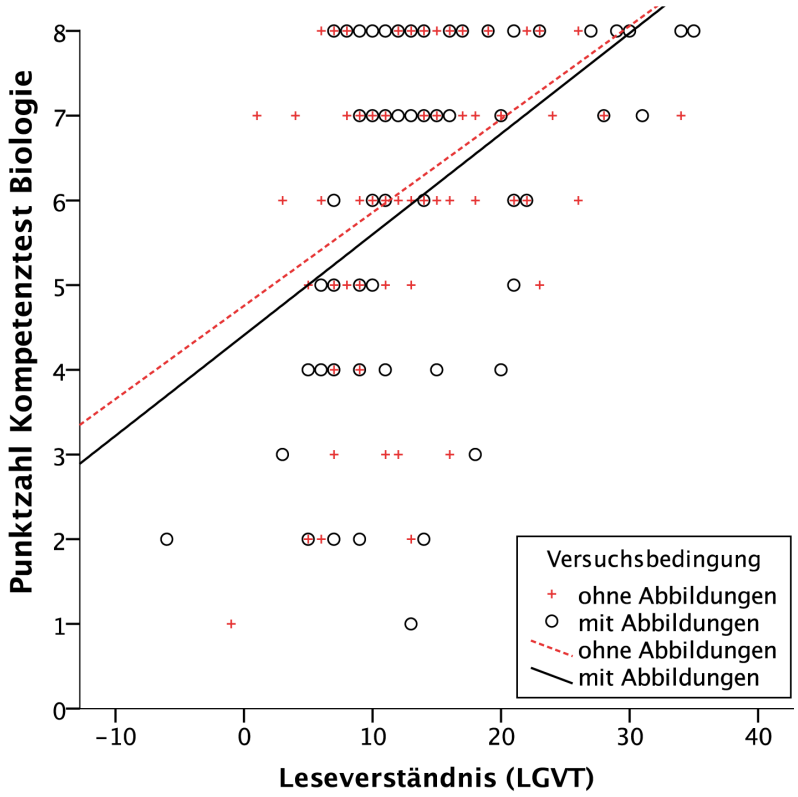


Abbildung 4.11: Streudiagramm mit Regressionsgeraden für den Zusammenhang zwischen der erzielten Punktzahl im Kompetenztest Biologie und dem Leseverständnis getrennt nach den beiden Versuchsbedingungen.

Auch hinsichtlich der Interaktion aus Lesegeschwindigkeit und Darbietungsformat wurde kein signifikanter Effekt gefunden (Tabelle 4.9 und Abbildung 4.12). Als signifikant erweisen sich wiederum nur die Haupteffekte für das Leseverständnis und die Lesegeschwindigkeit. Das Modell klärt 22 Prozent der Varianz auf.

Tabelle 4.9: Kovarianzanalyse für die Interaktion aus Lesegeschwindigkeit und Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_{p}
Leseverständnis	96.05	1	96.05	32.73	$< .001$	.18
Lesegeschwindigkeit	23.48	1	23.48	8.00	$< .01$	.05
Darbietungsformat	0.32	1	0.32	0.11	.742	.00
Lesegeschwindigkeit $\times$ Darbietungsformat	0.16	1	0.16	0.05	.818	.00
Fehler	425.48	145	2.93	—	—	—

**Anmerkung.* Sequentielle Varianzzerlegung. $R^2 = .220$.

4.3.5 Diskussion

Das Ziel, ein Instrument zu entwickeln, mit dem sich Kompetenzen im Bereich der Erkenntnisgewinnung im Fach Biologie in beiden Parallelformen reliabel testen lassen, konnte auch in der zweiten Vorstudie noch nicht gänzlich zufriedenstellend erreicht werden. Der Grund hierfür liegt vorrangig in der zu geringen Schwierigkeit mehrerer Items und infolgedessen der Skalen. Hiervon ist insbesondere die Skala *Text mit Abbildungen* betroffen. Aufgrund der Deckeneffekte werden die zu messenden Kompetenzen in Erkenntnisgewinnung im Fach Biologie noch nicht hinreichend zuverlässig erfasst. Ein weiteres Problem bei der Interpretation der Ergebnisse ergab sich infolge der Lesegeschwindigkeit in den beiden Gruppen, die in der Tendenz trotz Randomisierung einen schwachen Unterschied aufwies.

Die Skalenreliabilitäten der beiden Parallelformen liegen annähernd auf dem gleichen Niveau. Die in der ersten Vorstudie gefundenen erheblichen messtheo-

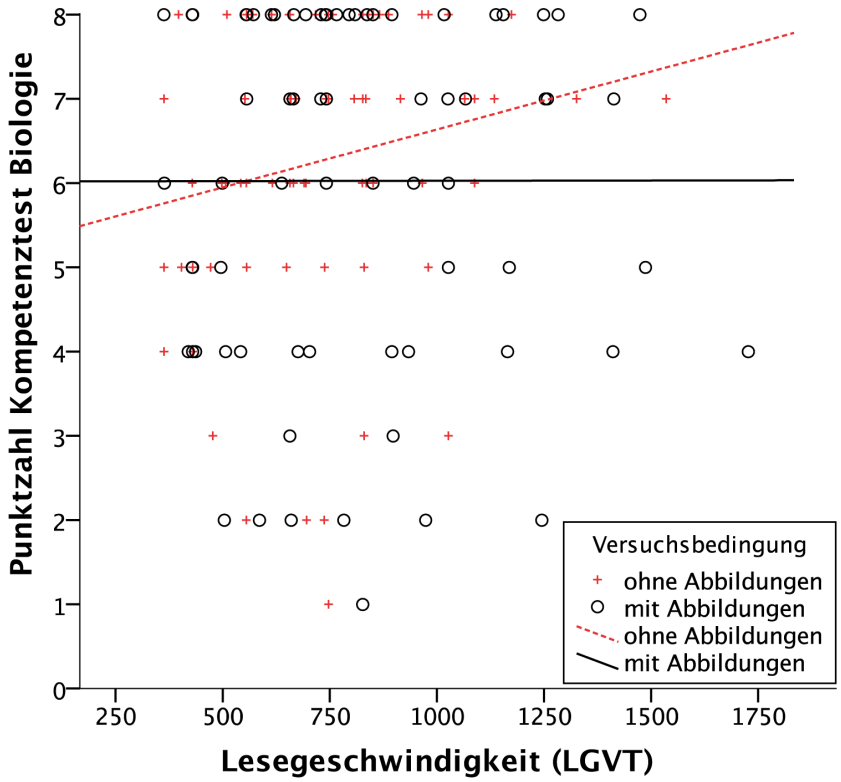


Abbildung 4.12: Streudiagramm mit Regressionsgeraden für den Zusammenhang zwischen der erzielten Punktzahl im Kompetenztest Biologie und der Lesegeschwindigkeit getrennt nach den beiden Versuchsbedingungen. Es deutet sich ein schwacher Effekt in der prognostizierten Richtung an, der sich aber als nicht signifikant erweist.

retischen Probleme der Skala *Text mit Abbildungen* konnten durch den Einsatz neuer Items deutlich reduziert werden. Das Entfernen kritischer Items führte auf beiden Skalen zu einer leichten Verbesserung der internen Konsistenz. Den Rückmeldungen der Testleiterinnen und Testleiter zufolge war die Bearbeitungszeit noch immer zu lang, sodass die Anzahl der Items weiter gesteigert werden könnte. Berechnungen mit der Spearman-Brown-Formel ergeben, dass sich Cronbach's α in der weniger reliablen Bedingung *Text mit Abbildungen* auf einen Wert von .74 verbessern ließe, wenn die Itemzahl noch einmal um den Faktor 1.5 (d. h. auf 18 Items) erhöht würde.

Neben den Anpassungen am Instrument wurden drei Hypothesenprüfungen durchgeführt. Es wurde keine hinreichende Evidenz zur Bestätigung der Hypothesen $H1_{Vorstudie2}$, $H2_{Vorstudie2}$ und $H3_{Vorstudie2}$ gefunden. Aufgrund der Deckeneffekte und der vergleichsweise geringen internen Konsistenz der Skala *Text ohne Abbildungen* sind die Ergebnisse der Analysen allerdings nur eingeschränkt interpretierbar. Bevor zuverlässige Aussagen über die Gültigkeit der aufgestellten Hypothesen möglich sind, bedarf es einer weiteren Optimierung des Instruments.

4.4 Hauptstudie

4.4.1 Ausgangssituation und Ziele der Studie

Die in diesem Kapitel vorgestellte Hauptstudie dient der abschließenden Prüfung der psychologischen Hypothesen dieser Arbeit. Hierzu gilt es, die noch verbliebenen messtheoretischen Probleme zu beseitigen, die sich in der zweiten Vorstudie abgezeichnet haben. Dies betrifft insbesondere die mangelnde Schwierigkeit einzelner Items und die daraus resultierenden Deckeneffekte auf den Skalen. Zur Lösung dieser Probleme kann bei der Aufgabenauswahl auf empirisch ermittelte Itemparameter aus der ESNaS-Pilotierungsstudie zurückgegriffen werden.

4.4.2 Zu prüfende psychologische Hypothesen

Die Hypothesen, die im Rahmen der Hauptstudie geprüft werden sollen, wurden bereits in Kapitel 2.5 eingeführt und lauten:

H1: Mit kombinierten Text-Bild-Aufgaben werden bessere Leistungen erzielt als mit reinen Textaufgaben.

H2: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Leseverständnis und Darbietungsformat feststellbar.

H3: Beim Zustandekommen der Testleistung in Erkenntnisgewinnung im Fach Biologie ist ein Interaktionseffekt aus Lesegeschwindigkeit und Darbietungsformat feststellbar.

H4: Personen mit niedrigem Vorwissen profitieren in höherem Maß vom Einsatz von Abbildungen als Personen mit hohem Vorwissen.

H5: Schülerinnen und Schüler, die mehr- bzw. fremdsprachig aufwachsen, profitieren in höherem Maß vom Einsatz von Abbildungen als solche, in deren Elternhäusern ausschließlich Deutsch gesprochen wird.

4.4.3 Methode

Variablen.

Die unabhängige Variable ist das Darbietungsformat der lösungsrelevanten Informationen in den Testaufgaben zur Erfassung von Kompetenzen im Fach Biologie (Kompetenzbereich *Erkenntnisgewinnung*). Die Variable kann zwei Ausprägungen annehmen: Text und statische Abbildungen. Da auch in der Hauptstudie nur die lösungsrelevanten Informationen, nicht aber die anderen Aufgabenbestandteile in Abbildungen überführt werden, ergeben sich die beiden Versuchsbedingungen *Text ohne Abbildungen* und *Text mit Abbildungen*. Als abhängige Variable werden die Leistungen im Biologie-Kompetenztest erfasst. Moderierende Effekte werden vom Leseverständnis und von der Lesegeschwindigkeit, von einem ggf. vorhandenen fremdsprachigen Elternhaus sowie

vom individuellen Vor- und Weltwissen der Schülerinnen und Schüler erwartet. Zum Zweck der konvergenten und diskriminanten Validierung sollen zudem die Schulnoten in den Fächern Biologie, Deutsch und Mathematik erhoben werden. Mögliche personenbezogene Störvariablen sind in Kapitel 4.1 theoretisch begründet und diskutiert.

Materialien.

Zum Zeitpunkt der Erstellung der Hauptstudie lagen der fachdidaktischen Leitung des Faches Biologie 45 Aufgabenstämme mit insgesamt 189 Testitems zum Kompetenzbereich *Erkenntnisgewinnung* aus dem ESNaS-Projekt vor. Alle Items sind in einer Pilotierungsstudie im Multimatrix-Design eingesetzt worden. Die Stichprobe dieser Pilotierung bestand aus $N = 2\,825$ Schülerinnen und Schülern, die zum Testzeitpunkt gerade das zehnte Schuljahr begonnen hatten. Es wurden Testungen in Gesamtschulen, Realschulen und Gymnasien in acht Bundesländern durchgeführt. Im Rahmen der Pilotierung wurden neben den Parametern der Rasch-Skalierung auch klassische Itemparameter zur Schwierigkeit und Trennschärfe ermittelt, die bei der Selektion geeigneter Aufgaben für die hier vorgestellte Studie berücksichtigt wurden. Ein Item wurde jeweils dann für den Einsatz als geeignet betrachtet, wenn es die folgenden Voraussetzungen erfüllte:

- getesteter Kompetenzteilbereich: *naturwissenschaftliche Untersuchungen* oder *naturwissenschaftliche Modellbildung* im Fach Biologie
- Antwortformat: geschlossene Antwort (Ankreuz-Aufgaben)
- a priori geschätzte Bearbeitungszeit: ≤ 2 Minuten
- klassische Itemschwierigkeit: $.20 \leq \ p_{\mathrm{i}} \leq .80$
- klassische Trennschärfe: $r_{\mathrm{it}} \geq .40$

Diese Kriterien wurden von 28 Items erfüllt. Eine Analyse der lösungsrelevanten Aufgabeninhalte ergab, dass sich die entsprechenden textbasierten Informationen im Fall von 11 Items nicht oder nur sehr schlecht in die Form von Abbildungen überführen ließen (es handelte sich um Gattungsbegriffe, logische Operatoren oder theoretische Konzepte). Sieben der verbleibenden 17

Items wiesen Inkompatibilitäten auf (die Aufgabenstämme enthielten Informationen, die für das Lösen anderer Items hilfreich gewesen wären). Somit eigneten sich 10 Items für den Einsatz in der Studie. Sie wurden um acht Items ergänzt, die nicht Bestandteil der ESNaS-Pilotierung waren, sich aber bereits in den Vorstudien der vorliegenden Arbeit bewährt hatten. Insgesamt standen somit 18 Items für die Erstellung der Testhefte zur Verfügung. Anhand von Berechnungen mit der Spearman-Brown-Formel wurde aus den Ergebnissen der zweiten Vorstudie gefolgert, dass die Anzahl von 18 Items zu einer zufriedenstellenden internen Konsistenz auf beiden Skalen führen müsste (vgl. Kapitel 4.3.5).

Die lösungsrelevanten Informationen der neu hinzugekommenen Items wurden entsprechend der in Kapitel 4.1 beschriebenen Vorgehensweise in Text bzw. Abbildungen überführt. Für die restlichen Items lagen aus den Vorstudien bereits jeweils zwei Formen vor. Sie wurden ohne Änderungen in die Hauptstudie übernommen.

Das Leseverständnis und die Lesegeschwindigkeit der Versuchspersonen wurde, wie bereits in den Vorstudien, mit dem LGVT 6–12 erfasst. Mögliche Moderatorvariablen wie die Sprache im Elternhaus, die Schulform und die Schulnoten in Biologie, Deutsch und Mathematik waren Bestandteil eines Personenfragebogens, der am Anfang aller Testhefte eingeheftet war (s. Anhang). Die Sprache im Elternhaus wurde im Gegensatz zur ersten Vorstudie nun über zwei Variablen erfasst (erste und ggf. zweite Sprache im Elternhaus), sodass zweisprachige Elternhäuser besser abgebildet werden konnten. Neben diesen Moderatorvariablen enthielt der Fragebogen noch die Variablen Alter und Geschlecht sowie einen persönlichen Code zur Zuordnung der Testhefte.

Räumlichkeiten und anwesende Personen.

Die Testung erfolgte in Klassenräumen der jeweils getesteten Schule. Die aus dem Unterricht bekannte Sitzordnung wurde—soweit ausreichend Platz zur Verfügung stand—so verändert, dass zwischen zwei Schülerinnen bzw. Schülern jeweils ein freier Platz lag. Neben den Versuchspersonen waren bei der Testung der Autor dieser Arbeit (als Versuchsleiter) sowie die jeweilige Lehrkraft für das Fach Biologie (als Aufsichtsperson) anwesend.

Stichprobe.

An der Hauptstudie nahmen 125 Schülerinnen und Schüler des 9. Schuljahres aus dem Bundesland Thüringen teil. Die getestete Schulform war die in Thüringen so bezeichnete Regelschule (eine landesspezifische Variante der Gesamtschule). Die von den Schülerinnen und Schülern angestrebten Schulabschlüsse waren der Hauptschulabschluss (14 Personen), der Realschulabschluss (84 Personen) und das Abitur (27 Personen). Der Anteil weiblicher Versuchspersonen betrug 49%; eine Person machte keine Angabe zum Geschlecht. Das Durchschnittsalter lag zum Testzeitpunkt bei 14.6 Jahren ($SD = 0.64$). In den Elternhäusern von 92 Personen wird ausschließlich Deutsch gesprochen; in 24 Fällen zusätzlich zum Deutschen eine weitere Sprache; die Elternhäuser von 8 Personen sind vollständig fremdsprachig. Die häufigsten Fremdsprachen sind Russisch (11 Personen) und Türkisch (10 Personen). Eine Person machte keine Angabe zur Sprache im Elternhaus.

Ablauf.

Der Untersuchungsablauf (Dauer, zeitliche Abfolge, Instruktionen und randomisierte Zuordnung der Testhefte) ist in Kapitel 4.1.6 beschrieben. Für den überwiegenden Teil der Schülerinnen und Schüler erwies sich die Testzeit für die Bearbeitung der 18 Items als ausreichend. Im Gegensatz zu den Vorstudien zeigte sich allerdings, dass ein Teil der Versuchspersonen zum Ende der Testzeit nicht mit der Bearbeitung aller Aufgaben fertig geworden war. Diesem Umstand wird bei der Auswertung durch zwei unterschiedliche Vorgehensweisen im Umgang mit fehlenden Werten Rechnung getragen (s. Kapitel 4.4.4).

4.4.4 Ergebnisse

Die Berechnungen zur Hauptstudie wurden mit den Programmen SPSS 20 für Mac und G*Power 3 für Mac (letzteres zur Bestimmung von Effektstärken) durchgeführt. Die Berechnung von Korrelationsvergleichen erfolgte gemäß Bortz (2005, S. 220f.) unter Anwendung einer von Field (2009) zur Verfügung gestellten SPSS-Syntax. Die Kodierung richtiger und falscher Antworten erfolgte den Richtlinien des ESNaS-Projekts gemäß. Dementsprechend wurde für richtig angekreuzte Antworten des Biologie-Tests jeweils ein Punkt ver-

geben; falsch angekreuzte und übersprungene Items wurden zunächst mit 0 Punkten gewertet.

Alternative Kodierung fehlender Werte.

Nicht alle Versuchspersonen haben die Testhefte in der vorgesehenen Zeit vollständig bearbeitet. Etwa ein Drittel der Versuchspersonen hat eine oder mehrere Aufgaben bei der Bearbeitung ausgelassen. Hiervon sind vor allem die jeweils letzten vier Items der Testhefte betroffen, die eine erhöhte Anzahl fehlender Werte aufweisen. In der Versuchsbedingung *Text ohne Abbildungen* wurden durchschnittlich 2.98 Aufgaben nicht bearbeitet ($SD = 4.01$), in der Bedingung *Text mit Abbildungen* 2.00 Aufgaben ($SD = 3.45$). Ein Vergleich der Verteilungen fehlender Werte in den beiden Versuchsbedingungen deutet auf einen schwachen Unterschied hin, der sich jedoch als nicht signifikant erweist (Tabelle 4.10). Um die Tatsache, dass die verfügbare Zeit für einige Versuchspersonen nicht ausreichte, dennoch zu berücksichtigen, wurden zusätzlich alternative Berechnungen vorgenommen, in denen übersprungene bzw. ausgelassene Aufgaben nicht als falsche Antworten, sondern als Missing Data gewertet wurden. Personen mit unvollständigen Datensätzen werden in dieser Bedingung aus den Berechnungen ausgeschlossen. Es verbleiben nur Personen, deren Gesamtpunktzahl im Biologie-Test nicht von der Bearbeitungsgeschwindigkeit abhängig war. Die Ergebnisse dieser abweichenden Vorgehensweise werden in den folgenden Abschnitten jeweils separat unter der Überschrift "Alternative Kodierung" berichtet. Die Skalenbildung erfolgt in beiden Fällen durch das Aufsummieren der richtigen Antworten.

Tabelle 4.10: Anzahl fehlender Werte in den beiden Versuchsbedingungen (deskriptive Werte, t-Test und Effektstärke)

Versuchsbedingung	n	M	SD	df	t_{emp}	p^*	d
ohne Abbildungen	63	2.98	4.01	123	1.47	.145	0.26
mit Abbildungen	62	2.00	3.45				

*zweiseitiger Test.

Test zur Erfassung biologischer Kompetenz in Erkenntnisgewinnung.

Die Testhefte ohne Abbildungen sind von 63 Personen bearbeitet worden; die Testhefte mit Abbildungen von 62 Personen. Für die Bedingung *Text ohne Abbildungen* ergibt sich ein Cronbach's α von .67. Die Trennschärfeparameter der Items können, abgesehen von vier Ausnahmen, als zufriedenstellend bezeichnet werden. Von den 18 Items weisen 17 eine mittlere Schwierigkeit ($.20 \leq p_{\mathrm{i}} \leq .80$) auf; das verbleibende Item verfehlt den als akzeptabel geltenden Bereich mit einer Schwierigkeit von .83 nur knapp.

In der Bedingung *Text mit Abbildungen* liegt Cronbach's α bei .73. Auch auf dieser Skala weisen vier Items eine mangelhafte Trennschärfe sowie ein Item (mit einer Lösungswahrscheinlichkeit von .81) eine tendenziell zu geringe Schwierigkeit auf. Durch das Entfernen der sechs kritischsten Items in beiden Bedingungen verbleiben 12 Items zur Bildung der Rohwertskalen. Hinsichtlich der internen Konsistenz werden damit zufriedenstellende Werte von $\alpha = .72$ (Skala *Text ohne Abbildungen*) bzw. $\alpha = .73$ (Skala *Text mit Abbildungen*) erreicht. Diese Skalen werden für alle weiteren Berechnungen herangezogen.

Die ursprünglich aus allen 18 Items gebildeten Rohwertskalen wiesen keine Decken- oder Bodeneffekte auf. Durch das Entfernen der kritischsten Items ergeben sich allerdings schwache Deckeneffekte (siehe Abbildungen 4.13 und 4.14). Die Verteilungsform der Histogramme weicht in beiden Versuchsbedingungen sichtbar von der Normalverteilung ab. Eine Überprüfung auf Normalverteilung ergibt mit dem K-S-Test allerdings keine statistisch bedeutsamen Abweichungen. Mit dem als strenger geltenden Verfahren nach Shapiro-Wilk kann eine signifikante Abweichung von der Normalverteilung nur für die Versuchsbedingung *Text mit Abbildungen* gefunden werden ($p < .05$), nicht jedoch für die andere Versuchsbedingung.

Alternative Kodierung. Werden unbearbeitete Aufgaben nicht als falsch, sondern als Missings gewertet, müssen 46 Personen aufgrund unvollständiger Datensätze aus den Berechnungen ausgeschlossen werden. Hierdurch reduziert sich die Anzahl der gültigen Fälle auf $N = 79$ (35 Personen in der Bedingung ohne Abbildungen und 44 in der Bedingung mit Abbildungen). Nach dem Entfernen kritischer Items (d. h. Items mit zu geringer Trennschärfe und / oder zu geringer bzw. zu hoher Itemschwierigkeit) verbleiben auch im Fall der alter-

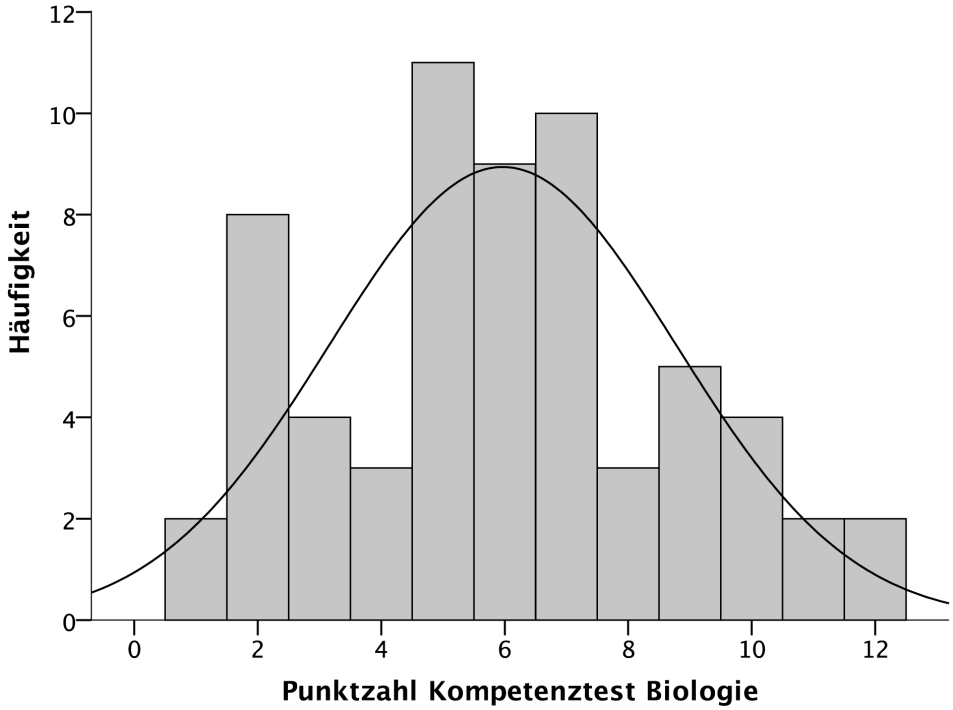


Abbildung 4.13: Histogramm für die Skala *naturwissenschaftiche Erkenntnisgewinnung in Biologie* in der Versuchsbedingung *Text ohne Abbildungen* (mit eingezeichneter Normalverteilungskurve).

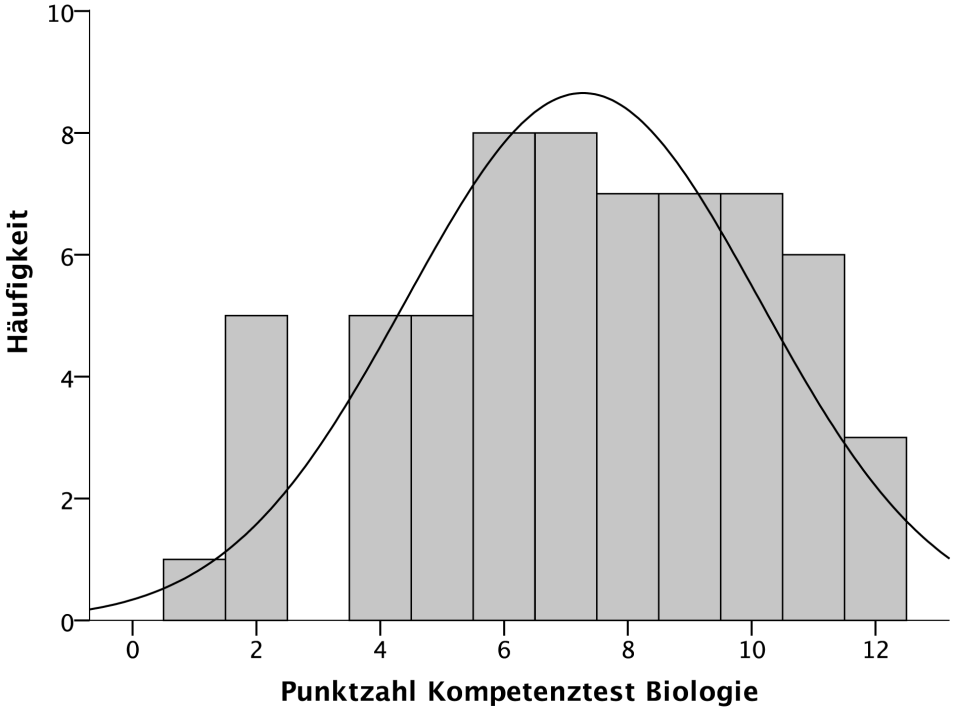


Abbildung 4.14: Histogramm für die Skala *naturwissenschaftiche Erkenntnisgewinnung in Biologie* in der Versuchsbedingung *Text mit Abbildungen* (mit eingezeichneter Normalverteilungskurve).

nativen Kodierung 12 Aufgaben zur Bildung der Rohwertskalen. Die interne Konsistenz der Skalen liegt mit Cronbach's $\alpha = .71$ (Skala *Text ohne Abbildungen*) bzw. Cronbach's $\alpha = .77$ (Skala *Text mit Abbildungen*) im akzeptablen Bereich.

Korrelationen mit Schulnoten.

Die aus 12 Items gebildeten Skalen des Biologie-Tests weisen in beiden Versuchsbedingungen signifikant von Null verschiedene, negative Korrelationen mit der Biologienote auf (das negative Vorzeichen resultiert aus der in Deutschland üblichen "umgekehrten" Rangfolge von Schulnoten). Gute Biologienoten gehen also mit hohen Punktzahlen im Test einher, schlechte Noten mit niedrigen Punktzahlen. Die Korrelationen mit der Mathematik- und der Deutschnote fallen jeweils nicht signifikant aus (Tabelle 4.11). Es fällt auf, dass der Zusammenhang mit der Biologienote in der Versuchsbedingung *Text mit Abbildungen* etwas stärker ist als in der Versuchsbedingung *Text ohne Abbildungen*. Der Korrelationsunterschied ist mit $p = .271$ nicht signifikant (einseitiger Test); die zugehörige Effektstärke beträgt $\varepsilon = 0.12$. Gleichzeitig reduziert sich der Zusammenhang zwischen den Leistungen im Kompetenztest und der Deutschnote durch den Einsatz von Abbildungen (der negative Korrelationskoeffizient kehrt sich in einen schwachen positiven Wert um). Der zugehörige Korrelationsunterschied weist eine mittlere Effektstärke von $\varepsilon = 0.29$ auf und verfehlt das Signifikanzniveau von fünf Prozent mit $p = .059$ (einseitig) nur knapp.

Tabelle 4.11: Korrelationen der Punktzahl in Biologie (Kompetenzbereich *Erkenntnisgewinnung*) mit Schulnoten getrennt nach Versuchsbedingung

	Biologienote	Mathematiknote	Deutschnote
Punktzahl Biologie (*Text ohne Abbildungen*)	$-.25^{*}$	.08	$-.17$
Punktzahl Biologie (*Text mit Abbildungen*)	$-.36^{**}$	$-.15$	.12

$^{*}p$ (einseitig) $< .05$. $^{**}p$ (einseitig) $< .01$.

Leistungsunterschiede nach angestrebtem Schulabschluss.

Obwohl die in der Hauptstudie getesteten Schülerinnen und Schüler alle dieselbe Schulform besuchen, unterscheiden sich ihre Leistungen in Abhängigkeit vom angestrebten Schulabschluss. Die Schülerinnen und Schüler, die den Hauptschulabschluss anstrebten, erzielten die niedrigste Punktzahl im Test, die Personen des Realschulzweigs schlossen besser ab und die Gymnasiasten erreichten die besten Ergebnisse (Tabelle 4.12).

Tabelle 4.12: Mittelwerte und Standardabweichungen der Punktzahl in Biologie (Kompetenzbereich *Erkenntnisgewinnung*) getrennt nach angestrebtem Schulabschluss und Versuchsbedingung

Versuchsbedingung	Hauptschule	Realschule	Abitur	gesamt
	3.33	5.63	8.14	5.97
Text ohne Abbildungen	(2.25)	(2.44)	(2.79)	(2.81)
	$n = 6$	$n = 43$	$n = 14$	$n = 63$
	5.13	7.22	8.77	7.27
Text mit Abbildungen	(3.18)	(2.56)	(2.86)	(2.86)
	$n = 8$	$n = 41$	$n = 13$	$n = 62$

Die Leistungen wurden auch auf eine mögliche Interaktion aus angestrebtem Schulabschluss und Darbietungsformat untersucht. Bei sequentieller Varianzzerlegung konnten sowohl für den angestrebten Schulabschluss ($F(2,119) = 12.16$, $p < .001$, $\eta^2_p = .17$) als auch für das Darbietungsformat ($F(1,119) = 9.04$, $p = .003$, $\eta^2_p = .07$) signifikante Haupteffekte nachgewiesen werden. Es besteht allerdings kein signifikanter Interaktionseffekt aus Darbietungsformat und Schulabschluss ($F(2,119) = 0.39$, $p = .677$, $\eta^2_p = .01$). Das Maß, in welchem die Regelschülerinnen und -schüler von den eingesetzten Abbildungen profitieren, ist in der getesteten Stichprobe also unabhängig vom angestrebten Abschluss (Abbildung 4.15).

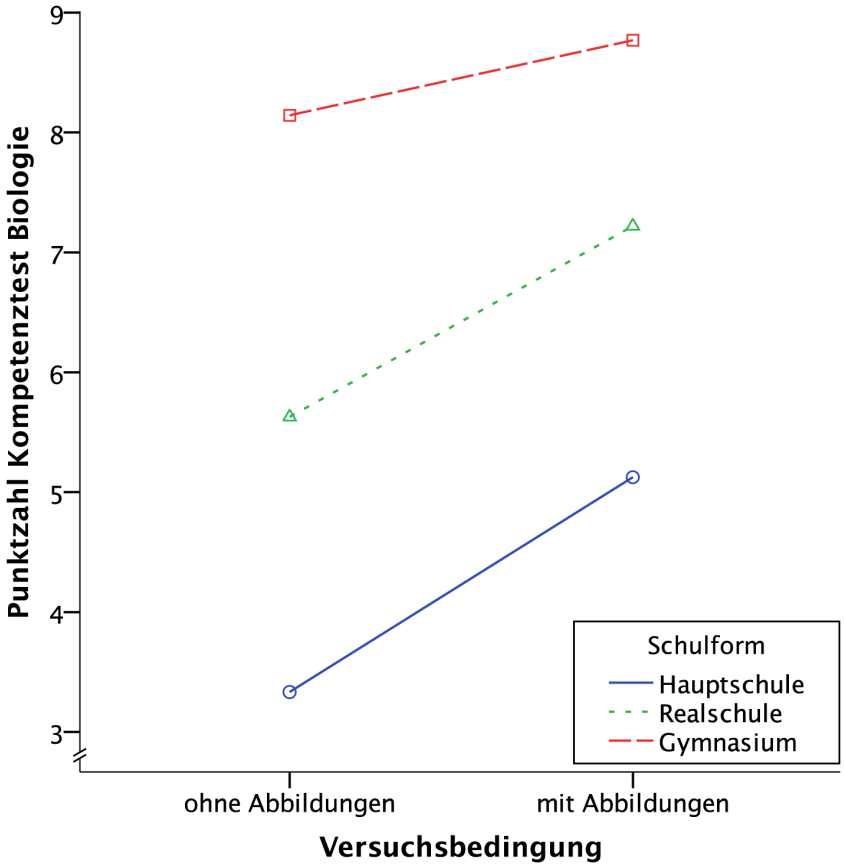


Abbildung 4.15: Haupteffekte für das Darbietungsformat und den angestrebten Schulabschluss. Eine Interaktion wurde nicht gefunden.

LGVT 6-12.

Das arithmetische Mittel für das Leseverständnis liegt bei 8.88 Rohwertpunkten (SD = 5.29) und die mittlere Lesegeschwindigkeit bei 620.98 Wörtern (SD = 182.61). Der Wert für die Lesegeschwindigkeit entspricht damit annähernd dem Normwert für die getestete Altersgruppe; der Wert für das Leseverständnis liegt hingegen etwas unter dem Normwert, der im LGVT-Manual publiziert ist.

Da die Personen den beiden Versuchsbedingungen randomisiert zugeordnet wurden, waren keine Gruppenunterschiede in den Leseleistungen zu erwarten. Um diese Annahme empirisch abzusichern, wurden die Ergebnisse des LGVT nach Gruppen getrennt ausgewertet und t–Tests für das Leseverständnis und die Lesegeschwindigkeit (mit dem Darbietungsformat des Biologie-Tests als Gruppenvariable) durchgeführt.

In der Versuchsbedingung *Text ohne Abbildungen* lag das Leseverständnis im Mittel bei 8.38 Rohwertpunkten (SD = 5.36); in der Bedingung *Text mit Abbildungen* wurden 9.39 Punkte erreicht (SD = 5.22). Die mittlere Lesegeschwindigkeit lag in der vierminütigen Testzeit in der Bedingung *Text ohne Abbildungen* bei 618.67 Wörtern (SD = 186.11) und in der Bedingung *Text mit Abbildungen* bei 623.34 Wörtern (SD = 180.47). Für keine der beiden Variablen wurden signifikante Unterschiede gefunden; die Ergebnisse liegen über dem für das Verwerfen der entsprechenden Nullhypothese üblichen Signifikanzniveau von .25 ($t_{\text{Leseverständnis}} = -1.1$, $p = .290$ [zweiseitig], $d = 0.19$; $t_{\text{Lesegeschwindigkeit}} = -0.14$, $p = .887$ [zweiseitig], $d = 0.03$).

Multimedia-Effekt (Hypothese H1).

Gemäß der Hypothese H1 wurde erwartet, dass kombinierte Text-Bild-Aufgaben leichter zu lösen sind als reine Textaufgaben ohne Abbildungen. Um diese Annahme zu überprüfen, wurden die in den beiden Parallelformen erzielten Leistungen mithilfe des t–Tests nach Student auf Mittelwertsunterschiede überprüft (Tabelle 4.13). Die Ergebnisse zeigen, dass die erzielten Leistungen in der Versuchsbedingung *Text mit Abbildungen* signifikant besser als in der Bedingung *Text ohne Abbildungen* sind. Wegen der leichten Deckeneffekte wurde eine zusätzliche Überprüfung mit einem nonparametrischen Verfahren

(U–Test nach Mann & Whitney) durchgeführt, welche zu dem selben Ergebnis kommt.

Tabelle 4.13: Vergleich der in den beiden Parallelformen erzielten Leistungen im Kompetenztest Biologie (deskriptive Werte, t-Test und Effektstärke)

Versuchsbedingung	*n*	*M*	*SD*	*df*	t_{emp}	p^*	*d*
ohne Abbildungen	63	5.97	2.81	123	–2.58	.006	0.46
mit Abbildungen	62	7.27	2.86				

*einseitiger Test.

Alternative Kodierung. Werden unbearbeitete Aufgaben alternativ nicht als falsch, sondern als Missings gewertet, fällt der Effekt schwächer aus und der Gruppenvergleich ergibt kein signifikantes Ergebnis, was rein rechnerisch sowohl auf die geringere Effektstärke ($d = 0.28$) als auch auf den kleineren Stichprobenumfang zurückgeführt werden kann (Tabelle 4.14).

Tabelle 4.14: Vergleich der in den beiden Parallelformen erzielten Leistungen im Kompetenztest Biologie bei alternativer Kodierung nicht bearbeiteter Aufgaben als Missings (deskriptive Werte, t-Test und Effektstärke)

Versuchsbedingung	*n*	*M*	*SD*	*df*	t_{emp}	p^*	*d*
ohne Abbildungen	35	6.89	2.87	77	–1.23	.111	0.28
mit Abbildungen	44	7.70	2.99				

*einseitiger Test.

Interaktionen aus Leseverständnis bzw. Lesegeschwindigkeit und Darbietungsformat (Hypothesen H2 und H3).

Gemäß den in Kapitel 2.4.2 beschriebenen theoretischen Annahmen wurde erwartet, dass der Effekt, den das Darbietungsformat auf die Leistungen im Biologie-Kompetenztest hat, eine Interaktion mit den Leseleistungen der Versuchspersonen aufweist. Die entsprechenden Hypothesen (H2 für die Interakti-

on mit dem Leseverständnis und H3 für die Interaktion mit der Lesegeschwindigkeit) wurden mithilfe von ATI-Analysen überprüft. Hierbei gingen die Werte der beiden LGVT-Skalen jeweils als Kovariaten in die varianzanalytischen Berechnungen ein.

Die Ergebnisse der Analyse zur Interaktion aus Leseverständnis und Darbietungsformat sind in Tabelle 4.15 zusammengefasst und in Abbildung 4.16 grafisch dargestellt. Es wurde ein starker und statistisch signifikanter Haupteffekt für das Leseverständnis gefunden ($F(1, 120) = 30.71$, $p < .001$, $\eta^2_p = .20$). Der Haupteffekt für das Darbietungsformat ist ebenfalls signifikant ($F(1, 120) = 5.46$, $p = .021$, $\eta^2_p = .04$). Für die Lesegeschwindigkeit wurde kein signifikanter Haupteffekt gefunden. Auch der vermutete Interaktionseffekt aus Leseverständnis und Darbietungsformat erweist sich als nicht signifikant. Die Hypothese H2 wird durch die Ergebnisse demnach nicht bestätigt.

Tabelle 4.15: Kovarianzanalyse für die Interaktion aus Leseverständnis und Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_p
Leseverständnis	203.27	1	203.27	30.71	$< .001$	.20
Lesegeschwindigkeit	5.84	1	5.84	0.88	.350	.01
Darbietungsformat	36.15	1	36.15	5.46	.021	.04
Leseverständnis $\times$ Darbietungsformat	1.89	1	1.89	0.29	.594	.00
Fehler	794.42	120	6.62	—	—	—

**Anmerkung.* Sequentielle Varianzzerlegung. $R^2 = .237$.

Für die Lesegeschwindigkeit und das Darbietungsformat wurde hingegen ein signifikanter ATI-Effekt gefunden ($F(1, 120) = 5.01$, $p = .027$, $\eta^2_p = .04$). Signifikante Haupteffekte ergaben sich erneut auch für das Leseverständnis ($F(1, 120) = 31.91$, $p < .001$, $\eta^2_p = .21$) und das Darbietungsformat ($F(1, 120) = 5.67$, $p = .019$, $\eta^2_p = .05$), wohingegen der Haupteffekt für die Lesegeschwindigkeit wiederum nicht signifikant ausfällt (Tabelle 4.16 und Abbildung 4.17). Die Hypothese H3 kann auf Basis der Ergebnisse angenommen werden.

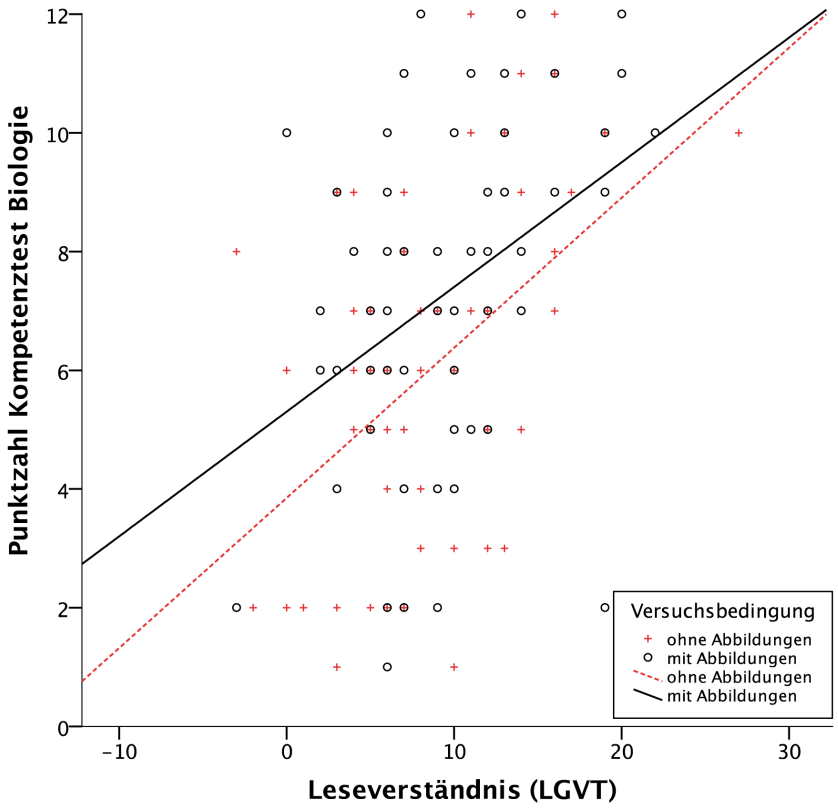


Abbildung 4.16: Streudiagramm mit Regressionsgeraden für den Zusammenhang zwischen der erzielten Punktzahl im Kompetenztest Biologie und dem Leseverständnis getrennt nach den beiden Versuchsbedingungen.

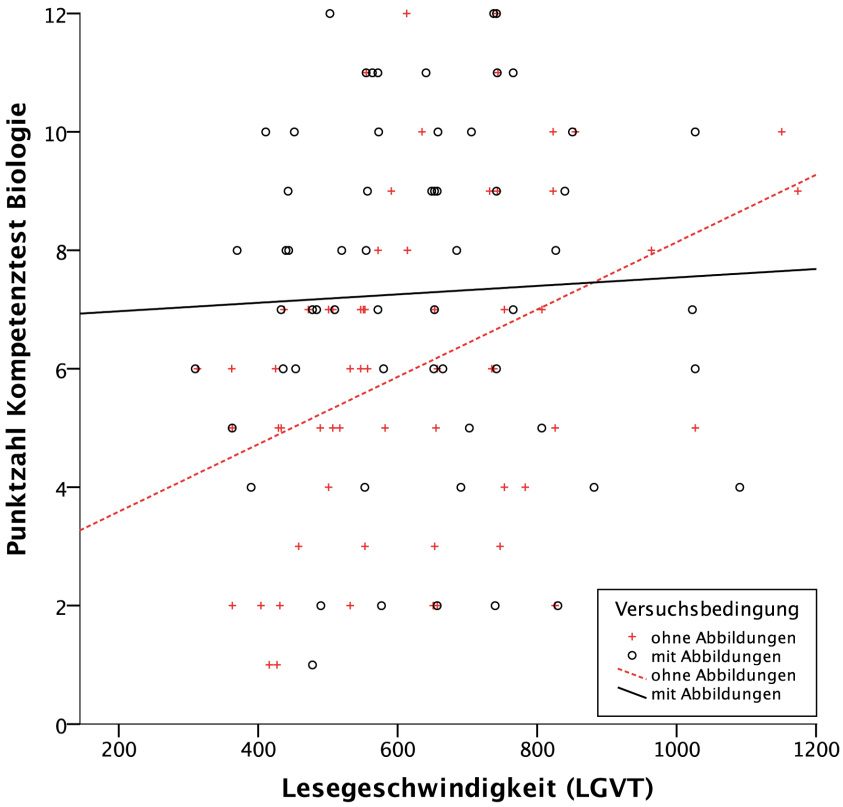


Abbildung 4.17: Streudiagramm mit Regressionsgeraden für den Zusammenhang zwischen der erzielten Punktzahl im Kompetenztest Biologie und der Lesegeschwindigkeit getrennt nach den beiden Versuchsbedingungen. Der Interaktionseffekt ist deutlich erkennbar.

Tabelle 4.16: Kovarianzanalyse für die Interaktion aus Lesegeschwindigkeit und Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_{p}
Leseverständnis	203.27	1	203.27	31.91	$< .001$	.21
Lesegeschwindigkeit	5.84	1	5.84	0.92	.340	.01
Darbietungsformat	36.15	1	36.15	5.67	.019	.05
Lesegeschwindigkeit × Darbietungsformat	31.89	1	31.89	5.01	.027	.04
Fehler	764.42	120	6.37	—	—	—

Anmerkung. Sequentielle Varianzzerlegung. $R^2 = .266$.

In Ergänzung zu den ATI-Analysen wurden die Korrelationen aus den Leistungen im Biologie-Test und den beiden LGVT-Skalen berechnet. Die Kovariate Leseverständnis korreliert mit der abhängigen Variablen (Punktzahl im Kompetenztest Biologie) zu .44 ($p < .001$). In der Versuchsbedingung *Text ohne Abbildungen* ist die Korrelation etwas höher ($r = .48$, $p < .001$) als in der Versuchsbedingung *Text mit Abbildungen* ($r = .38$, $p = .002$).

Die Korrelation aus der Kovariaten Lesegeschwindigkeit und der abhängigen Variablen (Punktzahl im Kompetenztest Biologie) beträgt .211 ($p = .018$). In der Versuchsbedingung *Text ohne Abbildungen* ist die Korrelation deutlich höher ($r = .38$, $p = .002$) als in der Versuchsbedingung *Text mit Abbildungen* ($r = .05$, $p = .728$).

Alternative Kodierung. Werden unbearbeitete Aufgaben nicht als falsch, sondern als Missings gewertet, ergeben sich für die ATI-Analysen nahezu identische Befunde. Es wird keine signifikante Interaktion aus Darbietungsformat und Leseverständnis gefunden ($F(1, 74) = 0.84$, $p = .361$, $\eta^2_{\text{p}} = .01$); die Haupteffekte für die Lesegeschwindigkeit und das Darbietungsformat fallen nicht signifikant aus. Hinsichtlich der Interaktion aus Darbietungsformat und Lesegeschwindigkeit ergibt sich auch im Fall der alternativen Kodierung ein signifikanter ATI-Effekt ($F(1, 74) = 7.34$, $p = .008$, $\eta^2_{\text{p}} = .09$); nicht signifikant werden die Haupteffekte für Lesegeschwindigkeit und Darbietungsformat.

Interaktion aus individuellem Vor- und Weltwissen und Darbietungsformat (Hypothese H4).

Gemäß der Hypothese H4 wurde erwartet, dass Schülerinnen und Schüler mit niedrigem Vor- und Weltwissen stärker vom Einsatz von Abbildungen profitieren als solche mit hohem Vor- und Weltwissen. Aus Gründen der Testökonomie wurde kein Wissenstest durchgeführt. Stattdessen wurde eine Einstufung anhand der Durchschnittsnote vorgenommen, die aus den letzten Zeugnisnoten in den Fächern Biologie, Mathematik und Deutsch gebildet wurde. Zunächst wurde mittels t-Test für unabhängige Stichproben abgesichert, dass keine systematischen Notenunterschiede zwischen den beiden Versuchsbedingungen bestehen. In der Bedingung *Text ohne Abbildungen* wurde eine Durchschnittsnote von 2.83 ermittelt ($SD = 0.65$); in der Bedingung *Text mit Abbildungen* liegt die Durchschnittsnote bei 2.73 ($SD = 0.63$). Es wurde kein signifikanter Unterschied gefunden; die Ergebnisse liegen deutlich über dem für das Verwerfen von Nullhypothesen üblichen Signifikanzniveau von .25 ($t = 0.72$, $p = .472$ [zweiseitig], $d = 0.14$).

Zur Unterteilung der Schülerinnen und Schüler in solche mit hohem und solche mit niedrigem Vorwissen wurde ein Mediansplit anhand der Durchschnittsnote durchgeführt. Das Vorwissen von Personen, deren Durchschnittsnote im Zahlenbereich unterhalb des Medians von 2.67 lag, wurde als "hoch" eingestuft (niedrige Noten bedeuten bessere Leistungen), das Vorwissen der anderen Schülerinnen und Schüler als "gering". Die Hypothese wurde mittels ANOVA (mit der Testleistung als abhängiger Variable sowie dem Darbietungsformat und dem Vorwissen als unabhängigen Variablen) überprüft. Wegen fehlender Werte für eine oder mehrere Noten wurden 14 Personen von den Berechnungen ausgeschlossen.

Personen mit niedrigem Vorwissen erreichten in der Bedingung *Text ohne Abbildungen* im Mittel 5.90 Punkte ($SD = 2.83$). In der Bedingung *Text mit Abbildungen* liegt der Mittelwert bei 7.15 Punkten ($SD = 2.75$). Von den Personen mit hohem Vorwissen wurden in der Bedingung *Text ohne Abbildungen* im Mittel 6.52 Punkte erreicht ($SD = 2.74$). In der Bedingung *Text mit Abbildungen* beträgt die mittlere Leistung 7.72 Punkte ($SD = 2.76$).

Die Ergebnisse der ANOVA sind in Tabelle 4.17 und in Abbildung 4.18 dargestellt. Signifikant wird nur der Haupteffekt für das Darbietungsformat. Der

Haupteffekt für das Vor- und Weltwissen ist nicht signifikant. Ein Interaktionseffekt wurde nicht gefunden. Die Daten liefern demnach keine Evidenz zur Bestätigung der Hypothese H4.

Tabelle 4.17: Varianzanalyse für die Interaktion aus dem individuellen Vor- und Weltwissen und dem Darbietungsformat

Varianzquelle	SS^{**}	df	MS	F_{emp}	p	η^2_{p}
Vor- und Weltwissen*	11.79	1	11.79	1.53	.218	.01
Darbietungsformat	41.98	1	41.98	5.46	.021	.05
Vor- und Weltwissen × Darbietungsformat	0.02	1	0.02	0.00	.961	.00
Fehler	822.61	107	7.69	—	—	—

Anmerkungen. *Mediansplit. **Sequentielle Varianzzerlegung. $R^2 = .061$.

Durch eine Dichotomisierung intervallskalierter Variablen—wie hier die Aufteilung in hohes vs. niedriges Vorwissen—ergibt sich immer ein gewisser Informationsverlust (Leutner & Rammsayer, 1995). Aus diesem Grund wurde in einer alternativen Form der Berechnung auf den Mediansplit verzichtet und die Durchschnittsnote als Kovariate bei sequentieller Varianzzerlegung in die Berechnung übernommen. Hierbei ergaben sich keine abweichenden Befunde; insbesondere die Interaktion aus Vorwissen und Darbietungsformat fällt auch hier nicht signifikant aus ($F(1, 107) = 1.34$, $p = .676$, $\eta^2_{\text{p}} = 0.00$).

Alternative Kodierung. Werden unbearbeitete Aufgaben nicht als falsch, sondern als Missings gewertet, ergeben sich für die oben berichteten Varianzanalysen weitgehend identische Befunde. Auch hier wird kein signifikanter Interaktionseffekt aus Darbietungsformat und Durchschnittsnote gefunden.

Interaktion mit der Sprache im Elternhaus (Hypothese H5).

Gemäß der Hypothese H5 wurde erwartet, dass Schülerinnen und Schüler aus fremd- oder mehrsprachigen Elternhäusern stärker von den eingesetzten Abbildungen profitieren müssten als solche, in deren Elternhaus ausschließlich Deutsch gesprochen wird. Die Sprachen im Elternhaus konnten anhand des

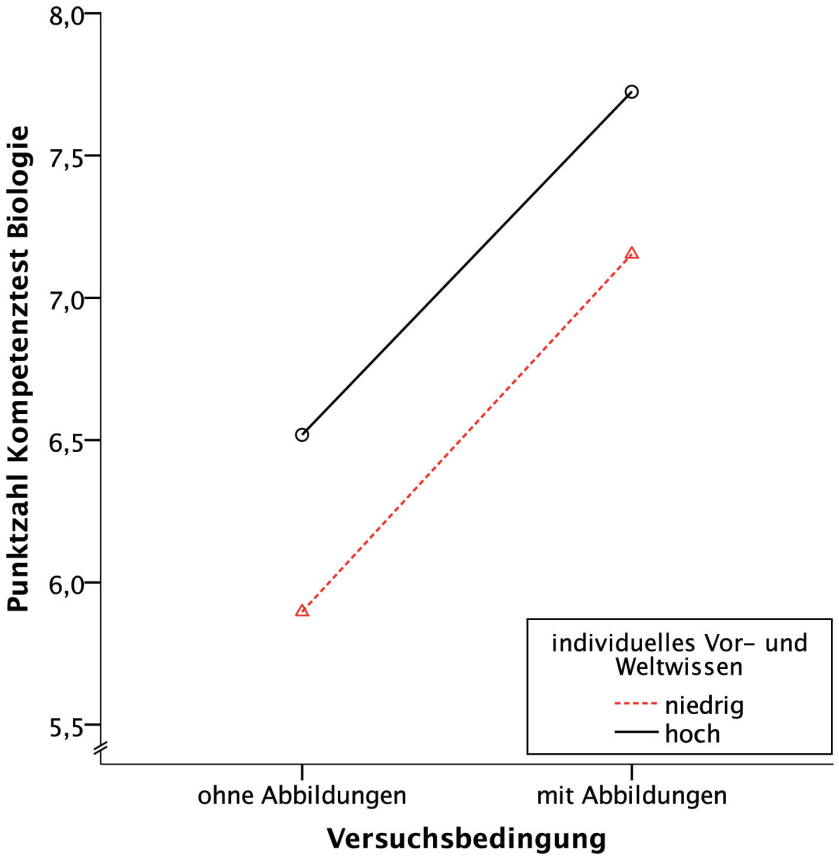


Abbildung 4.18: Punktzahl im Kompetenztest Biologie in Abhängigkeit vom Darbietungsformat und vom individuellen Vor- und Weltwissen. Es konnte ein signifikanter Haupteffekt für das Darbietungsformat nachgewiesen werden. Der Haupteffekt für das Vor- und Weltwissen erwies sich als nicht signifikant. Ein Interaktionseffekt wurde nicht gefunden.

eingesetzten Fragebogens in Erst- und Zweitsprache aufgeschlüsselt werden. Es bot sich an, die Versuchspersonen in drei Gruppen zu unterteilen: In der ersten Gruppe befinden sich Personen, in deren Elternhaus ausschließlich Deutsch gesprochen wird ($n = 92$). Diese Schülerinnen und Schüler erzielten in der Bedingung *Text ohne Abbildungen* eine mittlere Punktzahl von 6.35 Punkten ($SD = 2.83$) und in der Bedingung *Text mit Abbildungen* 7.47 Punkte ($SD = 2.72$). Die zweite Gruppe entstammt Haushalten, in denen neben dem Deutschen noch eine weitere Sprache gesprochen wird ($n = 24$). Die Leistungen dieser Personen fallen ähnlich aus wie in der ersten Gruppe; sie erzielten in der Bedingung *Text ohne Abbildungen* im Mittel 6.00 Punkte ($SD = 2.51$) und in der Bedingung *Text mit Abbildungen* 7.00 Punkte ($SD = 3.30$). Die dritte Gruppe besteht aus Schülerinnen und Schülern, die in vollständig fremdsprachigen Elternhäusern leben ($n = 8$). Diese Probanden erreichten in der Bedingung *Text ohne Abbildungen* im Mittel 3.17 Punkte ($SD = 1.94$) und in der Bedingung *Text mit Abbildungen* 6.50 Punkte ($SD = 3.54$).

Zur Überprüfung der Hypothese wurde zunächst eine ANOVA (mit der Testleistung als abhängiger sowie dem Darbietungsformat und der Fremdsprachigkeit im Elternhaus als unabhängiger Variablen) durchgeführt. Die Ergebnisse sind in Tabelle 4.18 und in Abbildung 4.19 dargestellt. Die Haupteffekte für die Sprache und das Darbietungsformat erwiesen sich jeweils als statistisch signifikant. Die grafische Darstellung zeigt, dass Personen aus deutschsprachigen Haushalten ungefähr in gleichem Maß von den eingesetzten Abbildungen profitierten wie Personen aus Elternhäusern, in denen zusätzlich zum Deutschen eine Fremdsprache gesprochen wird. Schülerinnen und Schüler, die in komplett fremdsprachigen Elternhäusern aufwachsen, profitieren anscheinend in deutlich stärkerem Maß von den Abbildungen. Ein entsprechender Interaktionseffekt erweist sich allerdings als nicht signifikant. Der Grund hierfür ist sicher auch in der Stichprobengröße der letzten Gruppe zu suchen, die mit nur acht Personen zu klein ist, um zuverlässige Aussagen zu ermöglichen.

Da die Teilstichprobe der Personen aus rein fremdsprachigen Elternhäusern mit nur acht Fällen sehr klein war, wurden alle Schülerinnen und Schüler, die entweder aus fremd- oder aus mehrsprachigen Elternhäusern stammen, zu einer Gruppe zusammengefasst und die ANOVA wiederholt. Hierbei ergab sich ebenfalls keine signifikante Interaktion ($F(2, 118) = 0.919$, $p = .340$, $\eta^2_{\mathrm{p}} = .01$).

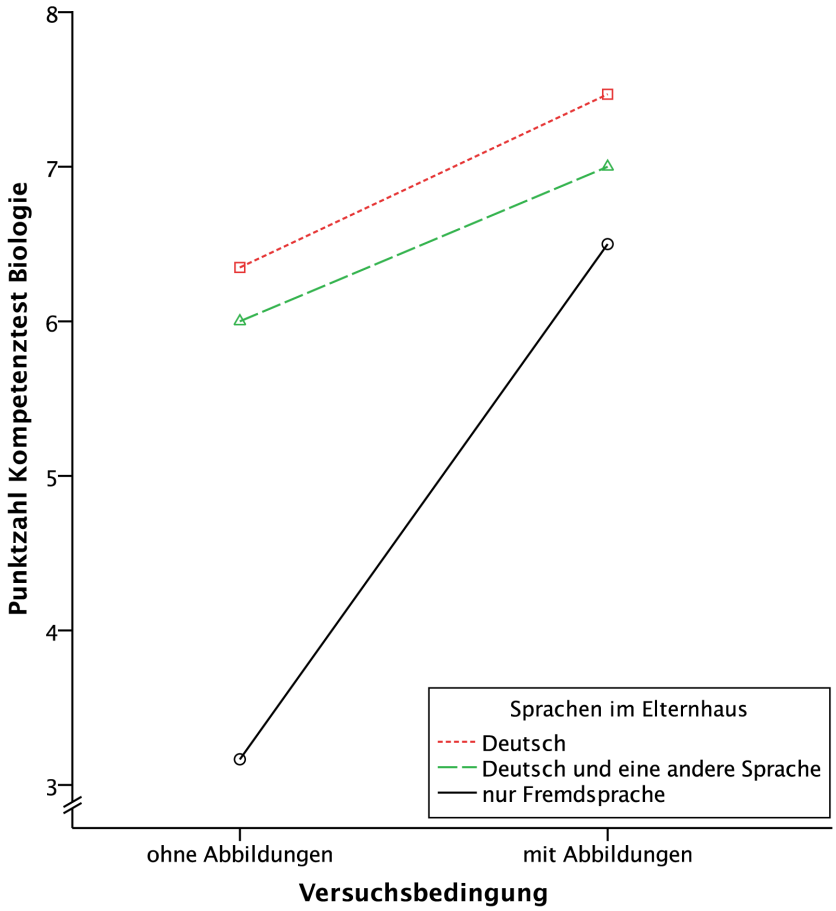


Abbildung 4.19: Haupt- und Interaktionseffekte für das Darbietungsformat und die Sprache im Elternhaus, hier getrennt nach Personen aus deutschprachigen, gemischtsprachigen und rein fremdsprachigen Elternhäusern. Aufgrund des zu geringen Umfangs der letztgenannten Gruppe ($n = 8$) kann der sich andeutende Interaktionseffekt nicht hinreichend zuverlässig interpretiert werden.

Tabelle 4.18: Varianzanalyse für die Interaktion aus der Sprache im Elternhaus und dem Darbietungsformat

Varianzquelle	SS^*	df	MS	F_{emp}	p	η^2_{p}
Sprache im Elternhaus	65.53	2	32.76	4.28	.016	.07
Darbietungsformat	44.05	1	44.05	5.75	.018	.05
Sprache im Elternhaus × Darbietungsformat	7.20	2	3.60	0.47	.626	.01
Fehler	903.31	118	7.66	—	—	—

** Anmerkung.* Sequentielle Varianzzerlegung. $R^2 = .114$.

Die Daten liefern demnach keine ausreichende Evidenz zur Bestätigung der Hypothese.

4.4.5 Diskussion

Ein Ziel der Hauptstudie bestand in der finalen Optimierung des Instruments, um die Hypothesen der vorliegenden Arbeit abschließend prüfen zu können. Dieses Vorhaben wurde erreicht. Die Auswahl von Items unter Berücksichtigung der Parameter aus der ESNaS-Pilotierungsstudie erwies sich als zielführend. Die interne Konsistenz der Skalen bewegt sich nun in beiden Versuchsbedingungen auf einem Niveau, das als zufriedenstellend bezeichnet werden kann.

Beide Skalen weisen signifikante Korrelationen mit der Biologienote, jedoch nicht mit der Mathematik- und der Deutschnote auf. Dies kann im Sinne einer curricularen Validierung als Indiz dafür gewertet werden, dass das Instrument Fähigkeiten misst, die im Biologieunterricht von Bedeutung sind, wohingegen mathematische und sprachliche Unterrichtsinhalte eine untergeordnete Rolle spielen. Auch die Tatsache, dass sich das erzielte Leistungsniveau in Abhängigkeit vom jeweils angestrebten Schulabschluss der Regelschülerinnen und Regelschüler unterscheidet, entspricht der Erwartung an einen Test, der unterrichtsbezogene Fähigkeiten misst.

Multimedia-Effekt.

Die Befunde zur Hypothese H1 bestätigen, dass sich die Schwierigkeit von Items zur Erfassung naturwissenschaftlicher Kompetenz durch den Einsatz lösungsrelevanter Abbildungen gegenüber reinen Textaufgaben signifikant reduzieren lässt. Hinsichtlich der Lösungswahrscheinlichkeit zeigt sich zwischen den beiden Versuchsbedingungen ein Unterschied mit einer mittleren Effektstärke ($d = 0.46$). Dieser illustriert den generellen Vorteil, den Abbildungen bei der Informationsverarbeitung in Testsituationen mit sich bringen. Der von Prenzel et al. (2002) gefundene Effekt, dass das Vorhandensein von Abbildungen in Testitems mit höheren Lösungswahrscheinlichkeiten einhergeht, konnte auf experimentellem Weg auf das Darbietungsformat der lösungsrelevanten Informationen zurückgeführt werden. Es kann davon ausgegangen werden, dass der Multimedia-Effekt, der sich beim multimedialen Lernen vielfach empirisch bewährt hat, auch in Testsituationen auftritt. Die theoretischen und fachdidaktischen Implikationen zu diesem Befund werden in Kapitel 5 abschließend diskutiert.

Konfundierung mit Leseleistungen.

Hinsichtlich der Rolle, die Leseleistungen beim Zustandekommen der Ergebnisse in Tests mit unterschiedlichen Darbietungsformaten spielen, ist die Befundlage uneindeutig. Während eine signifikante Interaktion aus Lesegeschwindigkeit und Darbietungsformat gefunden wurde, zeigte sich hinsichtlich des Leseverständnisses in der getesteten Stichprobe kein signifikanter Interaktionseffekt mit dem Darbietungsformat.

Wie stark die Konfundierung der Leistungen mit der Lesegeschwindigkeit durch den Einsatz von Abbildungen reduziert werden kann, lässt sich gut anhand der unterschiedlichen Korrelationen illustrieren, die in den beiden Versuchsbedingungen gefunden wurden. Die Korrelation aus der Punktzahl im Biologie-Kompetenztest und der Lesegeschwindigkeit, die in der Bedingung *Text ohne Abbildungen* .38 beträgt, konnte durch den Einsatz von Abbildungen auf .05 reduziert werden—hier besteht praktisch kein Zusammenhang zwischen der Lesegeschwindigkeit und den im Test erzielten Leistungen. Wenn Abbildungen zur Darbietung der lösungsrelevanten Informationen eingesetzt werden,

sind langsame Leserinnen und Leser beim Lösen der Aufgaben demnach nicht mehr benachteiligt. Der signifikante ATI-Effekt stützt diesen Befund.

Über die Konfundierung mit Leseverständnis kann auf Basis der getesteten Stichprobe hingegen keine eindeutige Aussage getroffen werden. Die Hypothese eines entsprechenden ATI-Effekts bestätigte sich nicht. Es deutete sich allerdings auch hier ein schwacher Effekt in der prognostizierten Richtung an. In sehr großen Stichproben (z. B. im Kontext von Large-Scale-Studien) könnte dieser Effekt demnach eine Rolle spielen.

Angesichts der vorgenannten Ergebnisse stellt sich die Frage, warum sich in der durchgeführten Untersuchung durch den Einsatz von Abbildungen zwar die Konfundierung der Testergebnisse mit der Lesegeschwindigkeit signifikant reduzieren ließ, nicht jedoch die Konfundierung mit dem Leseverständnis. Bei der inhaltlichen Interpretation dieses Befunds drängte sich zunächst der Verdacht auf, dass langsame Leserinnen und Leser vorrangig deshalb von Abbildungen profitieren, weil die Informationsentnahme aus den eingesetzten Grafiken weniger Zeit erfordert als die Entnahme derselben Informationen aus den entsprechenden Textpassagen in der Bedingung *Text ohne Abbildungen*. In diesem Fall könnten langsam lesende Schülerinnen und Schüler innerhalb der vorgegebenen Testzeit mehr Aufgaben bearbeiten, wodurch sie in der Versuchsbedingung mit Bildern weniger fehlende (d. h. im Fall der ursprünglichen Kodierung falsche) Lösungen zu verbuchen hätten. Tatsächlich deutete sich an, dass in der Bedingung *Text ohne Abbildungen* mehr Aufgaben unbearbeitet blieben als in der Bedingung *Text mit Abbildungen*. Dieser Unterschied erwies sich jedoch als nicht signifikant. Um einen möglichen Einfluss ganz auszuschließen, wurden die Daten zusätzlich in Form einer alternativen Kodierung fehlender Werte ausgewertet. Die Ergebnisse dieser Berechnungen zeigen, dass die ATI-Effekte in nahezu identischem Ausmaß auch dann auftreten, wenn sämtliche Datensätze mit fehlenden Werten aus den Analysen ausgeschlossen werden und somit nur jene Personen in die Berechnungen eingehen, die in der vorgegebenen Testzeit mit dem Bearbeiten aller Aufgaben fertig geworden sind. Damit konnte belegt werden, dass die Effekte nicht (bzw. nicht ausschließlich) auf eine erhöhte Anzahl an unbearbeiteten Aufgaben in der Versuchsbedingung *Text ohne Bilder* zurückzuführen sind. Dieser Erklärungsansatz musste also verworfen werden.

Eine weitere Möglichkeit, die unterschiedlich starken Effekte zu interpretie-

ren, kann aus den Theorien und Befunden zum Thema Lesegeschwindigkeit abgeleitet werden, die in Kapitel 2.3.4 dieser Arbeit vorgestellt wurden. Zahlreiche Autoren verweisen darauf, dass langsame Leserinnen und Leser zwar häufig in der Lage sind, Texte exakt zu dekodieren, sie dabei jedoch ein Mehr an Zeit und an kognitiven Kapazitäten benötigen (Artelt et al., 2007; Perfetti & Hogaboam, 1975; Schneider et al., 2007). Dies wird gelegentlich sogar zum Anlass genommen, die Lesegeschwindigkeit und nicht die Exaktheit der Dekodierung als das zuverlässigere Maß für die Unterscheidung zwischen starken und schwachen Leserinnen und Lesern heranzuziehen (Perfetti, 1985).

Durch den erhöhten kognitiven Aufwand langsam lesender Personen werden Ressourcen beansprucht, die eigentlich für hierarchiehöhere Verstehensprozesse benötigt würden (Perfetti, 1985). Infolgedessen wird die Kohärenzbildung auf Textebene beeinträchtigt (Artelt et al., 2007) und die aktive Sinnentnahme erschwert (Schneider et al., 2007). Es ist davon auszugehen, dass langsamen Leserinnen und Lesern somit auch weniger kognitive Ressourcen zur weiteren Verarbeitung der Aufgabeninhalte zur Verfügung stehen. Aus diesem Grund sind langsam lesende Schülerinnen und Schüler selbst dann beim Lösen der Aufgabenstellungen benachteiligt, wenn sie die Aufgabentexte fehlerfrei dekodieren können. Da die Verarbeitung von Bildern auf einem anderen Kanal erfolgt als die Verarbeitung von Texten (Schnotz, 2005), muss sich diese Benachteiligung langsamer Leserinnen und Leser eliminieren lassen, wenn die lösungsrelevanten Informationen in Form von Abbildungen präsentiert werden. Die Ergebnisse der ATI-Analyse deuten darauf hin, dass dies in der durchgeführten Studie gelungen ist. Um einzuschätzen, ob Abbildungen tatsächlich helfen, den Cognitive Load beim Bearbeiten der hier zum Einsatz gekommenen Aufgaben zu reduzieren, müsste diese Variable in einer zukünftigen Studie zusätzlich erhoben werden.

Auch motivationale Effekte müssen als mögliche Erklärung in Betracht gezogen werden. Es ist denkbar, dass die Schülerinnen und Schüler beim Bearbeiten der Aufgaben mehrmals zwischen der Fragestellung und dem Aufgabenstimulus hin und her wechseln, um die Informationen zu finden, die für die Lösung relevant sind. Hierbei sind langsame Leserinnen und Leser in der Text-Bedingung möglicherweise schneller von Frustration bzw. einem Nachlassen der Motivation betroffen als schnelle Leserinnen und Leser, sodass sie die

Suche abbrechen, bevor sie die relevanten Informationen gefunden haben. Es ist denkbar, dass dieser Effekt in der Bild-Bedingung nicht auftritt, weil das Auffinden der lösungsrelevanten Informationen kein erneutes Lesen, sondern nur das Betrachten von Abbildungen erfordert. Um diesen Erklärungsansatz zu überprüfen, müsste die Motivation bei einer Wiederholung der Studie zusätzlich erfasst werden.

Angesichts der uneindeutigen Befundlage für das Leseverständnis muss auch die Konstruktvalidität der zugehörigen LGVT-Skala noch einmal kritisch hinterfragt werden. Wie in Kapitel 4.1.4 beschrieben, erfasst der Test das Leseverständnis der Probanden nicht vorrangig im Sinne der Konstruktion einer kohärenten Repräsentation des Textes, sondern eher in Form der Präzision beim Lesen einzelner Wörter oder Wortgruppen. Personen, die vom gesamten Text eine durchaus adäquate Repräsentation gebildet haben, jedoch einzelne Details ungenau erfasst haben, werden im LGVT mit Punktabzug "bestraft". Genau dieses ungenaue Lesen wird durch die Aufgabenstellungen beim Bearbeiten der Biologie-Items aber möglicherweise verhindert, denn durch die Fragestellung wird der Fokus der Schülerinnen und Schüler gerade auf jene Textdetails gelenkt, die zum Lösen erforderlich sind. Aus diesem Grund sind Personen, die im LGVT niedrige Punktzahlen beim Leseverständnis aufweisen, bei der Lösung der ESNaS-Aufgaben möglicherweise nicht so stark benachteiligt, wie es bei Personen der Fall ist, die grundlegendere Probleme bei der Kohärenzbildung auf Textebene haben.

Dass eine niedrige Lesegeschwindigkeit, die als Indikator für mangelnde Kohärenzbildung und Probleme bei der Sinnentnahme gilt, tatsächlich nicht zwangsläufig mit einer geringen Punktzahl auf der LGVT-Skala *Leseverständnis* einhergeht, zeigen die Korrelationen, die zwischen diesen beiden Variablen festgestellt wurden. Die beiden Skalen des LGVT 6–12 weisen in der durchgeführten Hauptstudie nur einen Zusammenhang von .32 auf ($p < .001$). Es zeigt sich, dass ein nicht unerheblicher Teil der Schülerinnen und Schüler zwar langsam, aber dennoch präzise genug liest, um hohe Punktzahlen auf der Skala des Leseverständnisses zu erreichen. Gleichzeitig gibt es auch schnelle Leserinnen und Leser, die den Text offenbar ungenau gelesen haben und entsprechend nur eine niedrige Punktzahl beim Leseverständnis erreichen. Kombiniert man diese Informationen mit der Tatsache, dass die Texte der ESNaS-Aufgaben ge-

rade im Hinblick auf leseschwache Schülerinnen und Schüler möglichst einfach gehalten wurden, dann ist es wahrscheinlich, dass nicht ungenaue, sondern vor allem langsame Leserinnen und Leser bei der Bearbeitung reiner Textaufgaben benachteiligt sind.

Zusammenhang zwischen Konfundierung und Itemschwierigkeit.

Berücksichtigt man, dass die Lesegeschwindigkeit ein Indikator für den kognitiven Aufwand bei der Dekodierung, die Kohärenzbildung auf Textebene und somit für die Qualität des Lesevorgangs ist (Artelt et al., 2007; Perfetti, 1985; Perfetti & Hogaboam, 1975; Schneider et al., 2007), dann lässt sich die Interaktion aus der Lesegeschwindigkeit und dem Darbietungsformat auch heranziehen, um den gefundenen Multimedia-Effekt zu erklären. Abbildungen ermöglichen langsamen Leserinnen und Lesern eine bessere Kohärenzbildung auf Textebene bei gleichzeitig niedrigerem kognitiven Verarbeitungsaufwand. Dies führt zum einen dazu, dass die Aufgabentexte besser verstanden werden. Zum anderen stehen aber auch mehr kognitive Ressourcen für die eigentliche Bearbeitung der Aufgabe zur Verfügung. Beides zusammen erhöht die Wahrscheinlichkeit einer richtigen Lösung und reduziert den Anteil, den sprachliche Fähigkeiten beim Zustandekommen der Testergebnisse haben. Die von Prenzel et al. (2002) aufgestellte Annahme, dass Items mit Grafiken deshalb einfacher sind, weil das zum Lösen der Aufgabe erforderliche Textverständnis geringer ist, konnte mit den Ergebnissen der vorliegenden Arbeit zumindest teilweise experimentell bestätigt werden. Eine weitergehende Untersuchung dieses Erklärungsansatzes erfordert den gewonnenen Erkenntnissen zufolge die Verwendung alternativer Instrumente zur Erfassung des Leseverständnisses.

Parallel zu den Befunden hatte der Einsatz von Abbildungen auch einen schwachen Anstieg der Korrelation aus Testleistung und Biologienote sowie ein Absinken der Korrelation mit der Deutschnote zur Folge. Die Effekte sind zwar nicht signifikant, allerdings wird das fünfprozentige Signifikanzniveau im Fall der Deutschnote nur sehr knapp verfehlt. Dies kann als weiteres Indiz dafür gewertet werden, dass sich die curriculare Validität des Tests durch den Einsatz von Abbildungen leicht verbessert hat und dabei gleichzeitig die Bedeutung sprachlicher Fähigkeiten geringer ausfällt.

Interaktion mit dem Vor- und Weltwissen.

In der Hypothese H4 wurde prognostiziert, dass Personen mit geringem Vor- und Weltwissen stärker vom Einsatz von Abbildungen profitieren müssten als solche, die über umfangreicheres Wissen verfügen. Bei der Überprüfung der Hypothese wurde das Vorwissen aus testökonomischen Gründen in Form von Schulnoten erfasst.

Es ergab sich kein signifikanter Effekt. Ein naheliegender Grund hierfür könnte die Operationalisierung des Vorwissens in Form von Schulnoten sein. Diese ist möglicherweise nicht valide. So verweisen Ingenkamp und Lissmann (2008) darauf, "dass es weniger vom tatsächlichen Leistungsniveau, sondern stärker von der zufälligen Zugehörigkeit zu einer bestimmten Schulklasse abhängig ist, welche Zensuren ein Schüler erreicht" (S. 147). Zudem wurden verschiedene Schulformen bzw. –abschlüsse getestet, sodass dieselben Noten vermutlich nicht mit identischen Leistungen gleichzusetzen sind (die Vergabe der Note "gut" geht beispielsweise im Gymnasialzweig mit deutlich höheren fachlichen Anforderungen einher als im Hauptschulzweig).

Ein anderer möglicher Grund findet sich in einer Besonderheit der ESNaS-Aufgaben: Um den starken inhaltlichen Unterschieden der Lehrpläne in den teilnehmenden Bundesländern Rechnung zu tragen, wurde lösungsrelevantes fachliches Vorwissen bei den Schülerinnen und Schülern nicht als gegeben vorausgesetzt, sondern stattdessen in die Aufgabenstimuli integriert. Auf diese Weise sollte eine mögliche Konfundierung der Aufgaben zur Erfassung von Kompetenzen mit "reinem" Fachwissen verringert werden. Hierdurch reduziert sich eine Benachteiligung von Personen mit geringem fachlichen Vorwissen—und somit auch der in Hypothese H4 prognostizierte Effekt.

Um die Forschungsfrage nach der Rolle des Vorwissens abschließend zu klären, wäre eine weitere Untersuchung notwendig, in der (a) lösungsrelevante fachliche Informationen nicht in den Aufgabenstimuli enthalten sind und (b) die Variable *Vorwissen* in Form eines validen Tests erfasst wird. Diesem Ansatz sollte in zukünftigen Studien weiter nachgegangen werden.

Interaktion mit Sprache.

Bezüglich der Sprache deutete sich in den deskriptiven Daten an, dass Personen aus deutschsprachigen Elternhäusern in ähnlichem Maß von dem Einsatz von Abbildungen profitieren wie Schülerinnen und Schüler aus Elternhäusern, in denen neben dem Deutschen noch eine weitere Sprache gesprochen wird. Bei Personen aus rein fremdsprachigen Elternhäusern scheint der positive Effekt von Abbildungen deutlich stärker ausgeprägt zu sein als in den anderen beiden Gruppen. Die Gruppe der Schülerinnen und Schüler, in deren Elternhaus kein Deutsch gesprochen wird, war mit nur acht Personen allerdings zu klein, um sinnvoll interpretierbare Ergebnisse zu erhalten. Aufgrund der kleinen Stichprobe und der geringen Effektstärke konnte kein signifikanter Interaktionseffekt festgestellt werden, sodass keine ausreichende Evidenz zur Bestätigung der Hypothese H5 gegeben ist.

Zusätzlich wird ein stärkerer Effekt vermutlich auch durch die sprachliche Gestaltung der eingesetzten ESNaS-Aufgaben verhindert, die ein einwandfreies Verstehen der Aufgabentexte auch für Schülerinnen und Schüler mit eher geringem Sprachverständnis gewährleisten sollte. Anders ausgedrückt sind die Aufgabenstimuli der ESNaS-Items zumindest für Personen aus gemischtsprachigen Elternhäusern offenbar bereits verständlich genug formuliert, sodass ein Transfer einzelner Textabschnitte in Abbildungen nur noch eine geringfügige Verbesserung der Verstehensleistung mit sich bringt. Wenngleich somit keine ausreichende Evidenz für eine Bestätigung der Hypothese H5 gefunden wurde, so ist dieses Ergebnis doch im Sinne der Testfairness erfreulich.

Kapitel 5

Abschließende Diskussion

In der vorliegenden Dissertation wurde die Rolle von Leseleistungen beim Zustandekommen der Ergebnisse in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenzen untersucht. Zu Beginn der Arbeit wurden vier allgemeine Forschungsfragen formuliert:

1. Welche Rolle spielen individuelle Leseleistungen beim Zustandekommen der Ergebnisse in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenzen?

2. Müssen Leseleistungen im Kontext schriftlicher Instrumente zur Erfassung naturwissenschaftlicher Kompetenzen als konfundierende Variablen berücksichtigt werden—und falls ja: Wie stark sind die Konfundierungseffekte?

3. Welche Itemmerkmale sind ggf. für die Stärke der Konfundierungseffekte bedeutsam?

4. Welche Personenmerkmale sind ggf. für die Stärke der Konfundierungseffekte bedeutsam?

Um Antworten auf diese Fragen zu finden, wurde zunächst mithilfe einer Regressionsanalyse getestet, welchen Einfluss ausgewählte Aufgabenmerkmale auf die Itemparameter von Large-Scale-Aufgaben haben. Die Ergebnisse haben gezeigt, dass Aufgaben mit lösungsrelevanten Abbildungen im Vergleich zu reinen Textaufgaben signifikant geringere Itemschwierigkeiten aufweisen, also leichter zu lösen sind. Dies wurde darauf zurückgeführt, dass Abbildungen

möglicherweise helfen, die sprachlichen Anforderungen beim Lösen der Aufgaben zu reduzieren. Um diese Annahme unter kontrollierten Bedingungen empirisch abzusichern, wurden statische Abbildungen als alternatives Format zur Darbietung der lösungsrelevanten Informationen in den Aufgabenstimuli eines textbasierten Biologie-Tests eingesetzt und geprüft, wie sich diese Maßnahme (a) auf die erzielten Leistungen und (b) auf den Zusammenhang der Testergebnisse mit Leseverständnis und Lesegeschwindigkeit auswirkt.

Hierzu wurden aus dem Instrument zur Evaluation der Bildungsstandards im Fach Biologie Testitems zur Erfassung von Kompetenzen im Bereich der Erkenntnisgewinnung entnommen und in jeweils zwei Versionen umkonstruiert. Bei der ersten Version handelte es sich um reine Textaufgaben, in der zweiten Version waren die lösungsrelevanten Informationen in Form realitätsnaher Abbildungen dargestellt. Von diesem Unterschied abgesehen waren beide Varianten identisch. Im Rahmen von zwei Vorstudien wurden dann zunächst die Möglichkeiten zu einer adressatengerechten Visualisierung lösungsrelevanter Informationen ausgelotet und geeignete Items aus dem Itempool des Instruments zur Evaluation der Bildungsstandards ausgewählt. In der anschließenden Hauptstudie wurde das Instrument abschließend optimiert und die Prüfung der im Forschungsinteresse stehenden Hypothesen vorgenommen. Neben dem Leseverständnis und der Lesegeschwindigkeit der Probanden standen zwei weitere mutmaßlich moderierende Faktoren im Fokus: das individuelle Vor- und Weltwissen sowie die Sprache im Elternhaus.

Zwei der drei zentralen Hypothesen der Arbeit konnten im Experiment bestätigt werden. Durch den Einsatz lösungsrelevanter Abbildungen lässt sich die Schwierigkeit von textbasierten Testaufgaben generell reduzieren und eine Konfundierung mit Lesegeschwindigkeit signifikant verringern. Für die dritte Hypothese, in der eine Reduzierung der Konfundierung mit Leseverständnis vorhergesagt wurde, wurde ein schwacher Effekt in der prognostizierten Richtung gefunden, der auf Basis der getesteten Stichprobe jedoch nicht zu einem signifikanten Ergebnis führte und somit keine hinreichend belastbare Aussage über die Gültigkeit der Hypothese zulässt. In größeren Stichproben, wie sie beispielsweise in Large-Scale-Untersuchungen üblich sind, könnte der Effekt dennoch bedeutsam sein.

Hinsichtlich weiterer mutmaßlicher Moderatoreffekte ergab sich kein einheitliches Bild. In den deskriptiven Daten deutete sich zwar schwach an, dass bestimmte Personengruppen in höherem Maß von den eingesetzten Abbildungen profitieren als andere. Leichte Tendenzen zeigten sich beispielsweise für Hauptschülerinnen und Hauptschüler sowie für Personen aus rein fremdsprachigen Elternhäusern. Die entsprechenden Moderatoreffekte wurden jedoch in den jeweils getesteten Stichproben nicht signifikant. Dass keine ausreichende Evidenz für eine zuverlässige Bestätigung dieser Effekte gefunden wurde, konnte einerseits auf methodische Probleme, z. B. sehr kleine Teilstichproben oder die Operationalisierung des Vorwissens mittels Schulnoten, zurückgeführt werden. Andererseits zeigte sich, dass die Benachteiligung bestimmter Personengruppen in der Textversion der Aufgaben schwächer ausfiel als angenommen, sodass diese Personen in geringerem Maß vom Einsatz von Abbildungen profitierten, als es auf Basis der theoretischen Vorüberlegungen anzunehmen war. Im Hinblick auf das Ziel, die Bildungsstandards in den Naturwissenschaften mit einem möglichst fairen Instrument zu evaluieren, ist dies ein erfreulicher Befund.

5.1 Theoretische und fachdidaktische Implikationen

Das Ergebnis zur ersten Hypothese zeigt, dass der Multimedia-Effekt (Mayer, 1997), der sich in zahlreichen Studien zum Lernen mit Texten und Bildern empirisch bewährt hat, in ähnlicher Form auch beim Testen mit Text- und Text-Bild-Aufgaben gefunden werden kann. Der von Prenzel et al. (2002) berichtete Befund, dass Aufgaben mit Abbildungen eine geringere Schwierigkeit aufweisen als Aufgaben ohne Abbildungen, konnte auf Basis von zwei unterschiedlichen empirischen Ansätzen bestätigt werden und zeigte sich sowohl in der durchgeführten Regressionsanalyse auf Basis der ESNaS-Pilotierungsdaten als auch unter experimentellen Bedingungen. In beiden Fällen fiel die Lösungswahrscheinlichkeit signifikant höher aus, wenn die Aufgabenstimuli lösungsrelevante Informationen in Form von Abbildungen statt in Form von Text enthielten. In Abweichung zu den meisten Untersuchungen, die zum Multimedia-Effekt bisher publiziert sind, wurden die Bilder im Rahmen der experimentellen Studie

nicht in Ergänzung zum Text, sondern als Ersatz für bestimmte Textpassagen eingesetzt. Auf diese Weise wurde verhindert, dass die im Text und in den Abbildungen enthaltenen Informationen redundant waren. Dies war nötig, um die gefundenen Effekte zweifelsfrei auf das jeweils eingesetzte Darbietungsformat zurückführen zu können. Aus den Ergebnissen konnte geschlussfolgert werden, dass die verwendeten Abbildungen besser als Texte zur Bildung kohärenter mentaler Repräsentationen der lösungsrelevanten Aufgabeninhalte geeignet waren. Für die Entwicklerinnen und Entwickler zukünftiger Tests bedeutet das einerseits, dass den zu testenden Schülerinnen und Schülern das Verstehen der Aufgabeninhalte mithilfe von Abbildungen signifikant erleichtert werden kann. Andererseits muss die mit dem Einsatz von Abbildungen einhergehende Verringerung der Aufgabenschwierigkeiten insbesondere dann berücksichtigt werden, wenn ein bestimmtes Leistungsniveau durch Aufgaben mit geeigneter Schwierigkeit erfasst werden soll. Es ist theoretisch sogar möglich, die Schwierigkeit einer Aufgabe bei gleichbleibendem Inhalt allein durch die Wahl des einen oder anderen Formates gezielt zu manipulieren. Aufgrund des Einflusses, den das Darbietungsformat auf die Konfundierung mit Lesegeschwindigkeit hat, ist hiervon allerdings abzuraten.

Validität und Testfairness

Die Kritik, dass kompetenzorientierte Testinstrumente in den Naturwissenschaften in höchstem Maß mit Leseleistungen konfundiert sind (Klein, 2010; Rindermann, 2006), kann angesichts der gefundenen Ergebnisse zumindest für die ESNaS-Aufgaben nicht bestätigt werden. Die Korrelationen des Biologie-Kompetenztests mit Leseverständnis und Lesegeschwindigkeit bewegen sich in der Versuchsbedingung mit reinen Textaufgaben nur im mittleren Bereich. Es kann also davon ausgegangen werden, dass Leseverständnis und Lesegeschwindigkeit einen vergleichsweise moderaten Anteil beim Zustandekommen der Leistungen im ESNaS-Instrument haben. Die Befunde der Experimentalstudie haben zudem gezeigt, dass sich diese Konfundierungseffekte durch den Einsatz von Abbildungen noch weiter reduzieren lassen. Hiervon können insbesondere langsame Leserinnen und Leser profitieren. Durch die Überführung der lösungsrelevanten Informationen aus dem Textformat in statische Bilder konnte der Zusammenhang der Leistungen im Biologie-Kompetenztest mit Le-

segeschwindigkeit selbst in einer vergleichsweise kleinen Stichprobe signifikant reduziert werden. In der Versuchsbedingung *Text mit Abbildungen* sank der entsprechende Korrelationskoeffizient nahezu auf Null.

Hinsichtlich des Leseverständnisses wurde hingegen nur eine schwache Interaktion gefunden, die sich in der getesteten Stichprobe als nicht signifikant erwies. Für die Annahme, dass eine Konfundierung mit Leseverständnis durch den Einsatz von Abbildungen reduziert werden kann, wurde somit keine ausreichende Evidenz gefunden. Bei der Diskussion dieses Ergebnisses wurde unter anderem die Validität der LGVT-Skala "Leseverständnis" kritisch infrage gestellt. In zukünftigen Untersuchungen zur Prüfung der Hypothese sollte deshalb entweder vorab der Versuch einer Validierung dieser Skala unternommen oder ein alternatives Verfahren zur Erfassung des Leseverständnisses eingesetzt werden. Hierzu bieten sich andere ökonomische Instrumente (z. B. C–Tests) an. Daneben stehen je nach Umfang und Kontext der Studie unter Umständen auch bewährte Instrumente zur Verfügung, wie sie beispielsweise in PISA, PIRLS / IGLU oder bei der Evaluation der Bildungsstandards im Fach Deutsch zum Einsatz gekommen sind. Diese gehen aufgrund ihres Umfangs aber in der Regel mit einer längeren Testzeit als der LGVT oder die C–Tests einher.

Dass die Benachteiligung langsamer Leserinnen und Leser beim Bearbeiten von Testaufgaben mithilfe einfacher Abbildungen praktisch völlig eliminiert werden konnte, ist besonders im Sinne der Konstruktvalidität und der Testfairness ein erfreuliches und wünschenswertes Ergebnis. Gerade in großangelegten Studien wie der Evaluation der Bildungsstandards ist ein möglichst faires und valides Testen nicht nur von fachdidaktischem und psychometrischem, sondern auch von bildungspolitischem Interesse—insbesondere, was die Akzeptanz der eingesetzten Instrumente in Fachkreisen und in einer breiteren Öffentlichkeit angeht.

Die Vorteile, die Abbildungen beim Testen naturwissenschaftlicher Fähigkeiten bieten, sind mit dem Einsatz in Large-Scale-Assessments noch nicht erschöpft. Besonders der gefundene Multimedia-Effekt ist stark genug, um auch bei der Testung kleinerer Stichproben zu signifikanten Ergebnissen zu führen. Insofern erscheint es ratsam, den Effekt auch bei der Erfassung von Kompetenzen in kleineren Personengruppen nicht unberücksichtigt zu lassen—etwa bei Schuleignungstests oder in schulischen Leistungserhebungen wie Klausuren

und Prüfungen. Dabei gilt es selbstverständlich, den Aufwand zu berücksichtigen, der beim Anfertigen von Abbildungen entsteht. Dieser lässt sich allerdings—beispielsweise durch die Verwendung gemeinfreier oder unter offener Lizenz verfügbarer Abbildungen aus dem Internet—sehr gering halten.

5.2 Verallgemeinerbarkeit der Befunde, Limitationen und Ausblick

Zur Operationalisierung der naturwissenschaftlichen Kompetenzen von Schülerinnen und Schülern wurden Aufgaben verwendet, mit denen die Bildungsstandards im Bereich *Erkenntnisgewinnung* im Fach Biologie evaluiert werden. Es stellt sich die Frage, inwieweit sich die Befunde auf andere naturwissenschaftliche Fächer, auf andere Kompetenzbereiche und auch auf andere Instrumente zur Kompetenzmessung verallgemeinern lassen.

Hinsichtlich der Generalisierbarkeit über die Fächergrenzen hinweg kann davon ausgegangen werden, dass ähnliche Effekte wie die hier berichteten auch für die ESNaS-Aufgaben in Chemie und Physik nachweisbar sein müssten. Denn zum einen "werden Kompetenzen als Leistungsdispositionen betrachtet, die in einem gewissen Maß über ähnliche Situationen generalisierbar sind" (Hartig & Klieme, 2006, S. 129). Zum anderen kann eine gewisse Verallgemeinerbarkeit auch aufgrund des gemeinsamen Kompetenzmodells und der gemeinsamen Konstruktionsprinzipien angenommen werden, nach denen alle Aufgaben im Projekt ESNaS entwickelt wurden. Sowohl die Kompetenzteilbereiche als auch die Einteilung hinsichtlich der Merkmale *Komplexität* und *kognitive Prozesse* sind in allen drei Fächern identisch. Darüber hinaus ähneln sich die Aufgaben auch in Textlänge, sprachlichen Anforderungen und Antwortformaten. Aus denselben Gründen ist es auch wahrscheinlich, dass sich entsprechende Effekte nicht nur im Bereich *Erkenntnisgewinnung*, sondern auch in den Aufgaben zu den anderen bereits getesteten ESNaS-Kompetenzbereichen (*Umgang mit Fachwissen* und *Bewerten*) finden lassen.

Die Frage, ob sich die Befunde auch auf andere Papier-und-Bleistift-Tests zur Erfassung naturwissenschaftlicher Fähigkeiten oder sogar generell auf schriftliche Aufgaben zur Messung von Kompetenzen übertragen lassen, lässt sich weniger eindeutig beantworten. Die Stärke der Konfundierung mit Lese-

leistungen hängt vermutlich von einer ganzen Reihe von Aufgabenmerkmalen ab, die im Rahmen dieser Studie unberücksichtigt geblieben sind. Insbesondere bei sehr textlastigen Aufgabenformaten dürfte die Konfundierung mit Leseverständnis und Lesegeschwindigkeit höher ausfallen als in den hier verwendeten Aufgaben des ESNaS-Instruments, da in diesem Projekt eine zielgruppengerechte sprachliche Gestaltung durch verschiedene Maßnahmen gesichert wurde. Bei Aufgaben mit besonders langen oder komplizierten Texten erleichtern Abbildungen das Verstehen wahrscheinlich in höherem Maß als bei Aufgaben mit kurzen, einfachen Texten. Neben der Gestaltung des Aufgabenstimulus' wird auch das Antwortformat einen Einfluss auf die Konfundierung von Testergebnissen mit Leseleistungen bzw. mit sprachlichen Fähigkeiten haben. Offene Antwortformate erfordern nicht nur das Lesen der Aufgaben, sondern auch die Fähigkeit und Bereitschaft, die Lösungen eigenständig zu formulieren.

Aus den Grenzen, die bezüglich der Gestaltung von Texten und Abbildungen gesetzt sind, ergeben sich weitere Einschränkungen in der Verallgemeinerbarkeit der Ergebnisse. Wie in Kapitel 4.1.4 angemerkt, eignen sich bei weitem nicht alle Aufgabeninhalte gleichermaßen gut für eine Überführung in Abbildungen. Mit realitätsnahen Abbildungen, wie sie in der vorliegenden Studie zum Einsatz kamen, lassen sich zwar viele Objekte und Sachverhalte darstellen, die in der Realität mit bloßem Auge zu beobachten sind. Gattungsbezeichnungen, abstrakte Kategorien und logische Operatoren werden hingegen durch Begriffe ausgedrückt, die sich zwar einfach mit Sprache, jedoch nur schwer mit derartigen Bildern darstellen lassen. Zu einer grafischen Darstellung solcher Begriffe und Sachverhalte sind vermutlich abstraktere Abbildungstypen (z. B. Ikone oder Diagramme) geeignet, die aber bei den Rezipienten die Kenntnis entsprechender Konventionen voraussetzen. Ein typisches Beispiel aus dem Fach Biologie sind Blütendiagramme, die zwar einen Bezug zur Struktur, jedoch kaum zum tatsächlichen Aussehen echter Blüten aufweisen. Wenn den Versuchspersonen die zugrunde liegenden Darstellungskonventionen nicht bekannt sind, dann bieten solche Abbildungen keine Vorteile beim Lösen von Aufgaben, sondern erschweren die Lösung möglicherweise sogar.

Aber auch wenn sich die Inhalte von Aufgaben grundsätzlich für eine Visualisierung eignen, so ist damit noch längst nicht sichergestellt, dass den Rezipienten hierdurch tatsächlich das Verstehen der relevanten Informationen erleich-

tert wird. Vor dem Einsatz von Abbildungen in Testaufgaben sollte deshalb grundsätzlich überprüft werden, ob die Visualisierungen von der Zielgruppe als leicht verständlich wahrgenommen werden.

Mit Blick auf technologiebasierte Instrumente zur Kompetenzdiagnose bieten sich neben statischen Abbildungen weitere Möglichkeiten an, Informationen in anderen Formaten als geschriebenem Text zu präsentieren. Aus dem Integrierten Modell des Text-Bild-Verstehens (Schnotz, 2005) leitet sich beispielsweise gesprochener Text als ein weiteres geeignetes Darbietungsformat ab. Personen, die Probleme auf einer sehr basalen Ebene der Lesefähigkeit (z. B. beim Dekodieren von Buchstaben und Wörtern) haben, profitieren möglicherweise, wenn sie Aufgabentexte nicht in schriftlicher, sondern in gesprochener Form dargeboten bekommen. Daneben bieten computerbasierte Tests die Möglichkeit, bewegte Bilder einzusetzen. Auch diese haben sich—sowohl alleine als auch in Kombination mit gesprochenen Texten—bereits als förderlich für das Verstehen bestimmter Sachverhalte erwiesen (Hartmann, 2006; Höffler, 2007) und eröffnen somit weitere Möglichkeiten für die Gestaltung valider und fairer Testaufgaben.

Literatur

Artelt, C., Drechsel, B., Bos, W. & Stubbe, T. B. (2008). Lesekompetenz in PISA und PIRLS/IGLU – ein Vergleich. In M. Prenzel and J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006* (S. 35–52). Wiesbaden: Verlag für Sozialwissenschaften.

Baker, E. L., O'Neil, H. F. & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210–1218.

Baumert, J., Brunner, M., Lüdtke, O. & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? – Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. *Psychologische Rundschau, 58*, 118–145.

BMBF [Bundesministerium für Bildung und Forschung] (Hrsg.) (2007a). *Förderung von Lesekompetenz. Expertise.* Bonn: BMBF.

BMBF [Bundesministerium für Bildung und Forschung] (Hrsg.) (2007b). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bonn: BMBF.

Bonsen, M., Kummer, N. & Bos, W. (2008). Schülerinnen und Schüler mit Migrationshintergrund. In W. Bos, M. Bonsen, J. Baumert, M. Prenzel, C. Selter & G. Walther (Hrsg.), *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 157–175). Münster: Waxmann.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler.* Heidelberg: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Heidelberg: Springer.

Bos, W., Bonsen, M., Baumert, J., Prenzel, M., Selter, C. & Walther, G. (Hrsg.) (2008). *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen*

Vergleich. Münster: Waxmann.

Bransford, J. D., Barclay, J. R. & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology, 3*, 193–209.

Bransford, J. D. & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717–726.

Brünken, R. & Leutner, D. (2005). Individuelle Unterschiede beim Lernen mit neuen Medien - neue Wege in der ATI-Forschung? In S. R. Schilling, J. R. Sparfeldt & C. Pruisken (Hrsg.), *Aktuelle Aspekte pädagogisch-psychologischer Forschung. Detlef H. Rost zum 60. Geburtstag* (S. 25–40). Münster: Waxmann.

Brünken, R., Seufert, T. & Zander, S. (2005). Förderung der Kohärenzbildung beim Lernen mit multiplen Repräsentationen. *Zeitschrift für Pädagogische Psychologie, 19*, 61–75.

Bybee, R. W. (2002). Teaching science as inquiry. In J. Minstrell & E. H. van Zee (Hrsg.), *Inquiring into inquiry learning and teaching in science* (S. 20–46). Washington: American Association for the Advancement of Science.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.

Cronbach, L. J. & Snow, R. E. (1969). *Individual differences in learning ability as a function of instructional variables. Final report* (USOE contract OEC-4-6-061269-1217). Stanford: Stanford University, California School of Education. Zugriff am 05.03.2012. Verfügbar unter http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED029001

Farcot, M. & Latour, T. (2009). Transitioning to computer-based assessments: A question of costs. In F. Scheuermann & J. Björnsson (Hrsg.), *The transition to computer-based-assessment. New approaches to skills assessment and inplications for large-scale testing* (S. 108–116). Luxembourg: Office for Official Publications of the European Communities.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Field, A. (2009). *Discovering statistics using SPSS.* London: Sage.

Grube, C. (2010). *Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung. Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I.* Elektronische Dissertation, Universität Kassel.

Grube, C., Hartmann, S. & Mayer, J. (2008). Kompetenzstrukturmodell zum wissenschaftlichen Denken. In Leibniz-Institut für die Pädagogik der Naturwissenschaften (Hrsg.), *Kompetenz. Modellierung, Diagnostik, Entwicklung, Förderung. 71. Tagung der AEPF in Kiel* (S. 89). Kiel: IPN.

Grube, C., Möller, A. & Mayer, J. (2007). Dimensionen eines Kompetenzstrukturmodells zum Experimentieren. In H. Bayrhuber, F. X. Bogner, D. Graf, H. Gropengießer, M. Hammann, U. Harms et al. (Hrsg.), *Ausbildung und Professionalisierung von Lehrkräften. Internationale Tagung der Fachgruppe Biologiedidaktik im VBiO* (S. 31–34). Kassel: Universität Kassel.

Haarmann, H. (1991). *Universalgeschichte der Schrift.* Frankfurt: Campus Verlag.

Haldane, S. (2009). Delivery platforms for national and international computer-based surveys. History, issues and current status. In F. Scheuermann & J. Björnsson (Hrsg.), *The transition to computer-based-assessment. New approaches to skills assessment and implications for large-scale testing* (S. 63–67). Luxembourg: Office for Official Publications of the European Communities.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 83–99). Weinheim: Beltz.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer.

Hartmann, S. (2006). *Statisches vs. bewegtes Bild. Eine empirische Studie zur Gestaltung multimedialer Lernumgebungen.* Unveröffentlichte Magisterarbeit, Universität Erfurt.

Helgeson, S. L. (1993). Research on problem solving: Middle school. In D.

Gabel (Hrsg.), *Handbook of research on science teaching and learning* (S. 248–268). New York: MacMillan.

Hemminger, U., Roth, E., Schneck, S., Jans, T. & Warnke, A. (2000). Testdiagnostische Verfahren zur Überprüfung der Fertigkeiten im Lesen, Rechtschreiben und Rechnen. Eine kritische Übersicht. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 28*, 188–201.

Höffler, T. N. (2007). *Lernen mit dynamischen Visualisierungen. Metaanalyse und experimentelle Untersuchungen zu einem naturwissenschaftlichen Lerninhalt.* Elektronische Dissertation, Universität Duisburg-Essen.

Hollingworth, L., Beard, J. J. & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. *Practical Assessment, Research & Evaluation, 12*(18). Zugriff am 05.03.2012. Verfügbar unter http://pareonline.net/getvn.asp?v=12&n=18

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik.* Weinheim: Beltz.

Jude, N. & Wirth, J. (2007). Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 49–56). Berlin: Bundesministerium für Bildung und Forschung.

Kampa, Nele (2010, July). *Dimensionality of competence in biology.* Paper presented at the meeting of the EARLI Junior Researchers, Frankfurt/Main, Germany.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 135–153.

Klammer, B. (2005). *Empirische Sozialforschung. Eine Einführung für Kommunikationswissenschaftler und Journalisten.* Konstanz: UVK.

Klein, H. P. (2010). Die neue Kompetenzorientierung: Exzellenz oder Nivellierung? *Zeitschrift für Didaktik der Biowissenschaften, 1*, 15–26.

Klein-Braley, C. (1985). A cloze-up on the C-Test. A study in the construct validation of authentic tests. *Language Testing, 2*, 76–104.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschrei-

bung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik, 52*, 876–903.

Köller, O. (2007). Bildungsstandards, einheitliche Prüfungsanforderungen und Qualitätssicherung in der Sekundarstufe II. In D. Benner (Hrsg.), *Bildungsstandards. Chancen und Grenzen, Beispiele und Perspektiven* (S. 13–28). Paderborn: Schöningh.

Kozma, R. B. (2000). The use of multiple representations and the social construction of understanding in chemistry. In M. Jacobson & R. Kozma (Hrsg.), *Innovations in science and mathematics education: Advanced designs for technologies of learning* (S. 11–46). Mahwah: Erlbaum.

Kürschner, C. & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. *Psychologische Rundschau, 59*, 139–149.

Landerl, K. (2001). Basale Lesefertigkeiten – Reading Speed. In G. Haider (Hrsg.), *PISA 2000 – Technischer Report: Ziele, Methoden und Stichproben des österreichischen PISA-Projekts* (S. 112–116). Innsbruck: Studien Verlag.

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65–99.

Leutner, D., Klieme, E., Meyer, K. & Wirth, J. (2004). Problemlösen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 147–175). Münster: Waxmann.

Leutner, D. & Rammsayer, T. (1995). Complex trait-treatment-interaction analysis: A powerful approach for analysing individual differences in experimental designs. *Personality and Individual Differences, 19*, 493–511.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse.* Weinheim: Belz.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for english language learners in math tests. *Educational Assessment, 14*, 160–179.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidakti-*

schen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden (S. 177–186). Berlin: Springer.

Mayer, J., Harms, U., Hammann, M., Bayrhuber, H. & Kattmann, U. (2004). Kerncurriculum Biologie der gymnasialen Oberstufe. *Der mathematische und naturwissenschaftliche Unterricht, 57*, 166–173.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*, 1–19.

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*, 125–139.

Mayer, R. E. (Hrsg.) (2005a). *The Cambridge handbook of multimedia learning.* Cambridge: Cambridge University Press.

Mayer, R. E. (2005b). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer (Hrsg.), *The Cambridge handbook of multimedia learning* (S. 183–200). Cambridge: Cambridge University Press.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory.* Amsterdam: North-Holland.

Moosbrugger, H. (2008a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 215–259). Heidelberg: Springer.

Moosbrugger, H. (2008b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 99–112). Heidelberg: Springer.

National Research Council (2000). *Testing english-language learners in U.S. schools. Report and workshop summary.* Washington, D.C.: National Academy Press.

Olson, J. F., Martin, M. O. & Mullis, I. V. S. (Hrsg.) (2008). *TIMSS 2007 technical report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Organisation for Economic Cooperation and Development (Hrsg.) (2003). *Literacy skills for the world of tomorrow - Further results from PISA 2000.* Paris: OECD.

Organisation for Economic Cooperation and Development (Hrsg.) (2009). *PISA 2006 technical report.* Paris: OECD.

Paetsch, J. & Radmann, S. (2011). Einfluss des Leseverständnisses auf die Mathematikleistungen von Kindern deutscher und nichtdeutscher Herkunftssprache [Abstract]. In M. Mitchell & J. Abel (Hrsg.), *Abstractband der Sektionstagung für empirische Bildungsforschung 2011. Nationale und regionale empirische Bildungsforschung* (S. 116). Bamberg: Universität Bamberg.

Peirce, C. S. (1906). Prolegomena to an apology for pragmaticism. *The Monist, 16*, 492–546.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Perfetti, C. A. & Hogaboam, T. W. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67,* 461–469.

PISA–Konsortium Deutschland (Hrsg.) (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs.* Münster: Waxmann.

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

Popper, K. R. (1994). *Logik der Forschung.* Tübingen: Mohr.

Popper, K. R. & Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature, 302*, 687–688.

Prenzel, M. & Baumert, J. (Hrsg.) (2008). Vertiefende Analysen zu PISA 2006. *Zeitschrift für Erziehungswissenschaft, Sonderheft 10.*

Prenzel, M., Drechsel, B., Carstensen, C. H. & Ramm, G. (2004). PISA 2003 – eine Einführung. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 13–46). Münster: Waxmann.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30*, 120–135.

Prenzel, M., Walter, O. & Frey, A. (2007). PISA misst Kompetenzen. Eine Replik auf Rindermann (2006): Was messen internationale Schulleistungsstudien? *Psychologische Rundschau, 58*, 128–136.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? *Psychologische Rundschau, 57*, 69–86.

Ropohl, M. (2010). *Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. Entwicklung und Analyse von Testaufgaben.* Berlin: Logos.

Rost, D. H. (2005). *Interpretation und Bewertung pädagogisch-psychologischer Studien. Eine Einführung.* Weinheim: Beltz.

Scheuermann, F. & Björnsson, J. (Hrsg.) (2009). *The transition to computer-based-assessment. New approaches to skills assessment and implications for large-scale testing.* Luxembourg: Office for Official Publications of the European Communities.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12. Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12.* Göttingen: Hogrefe.

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review, 14*, 101–120.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Hrsg.), *The Cambridge handbook of multimedia learning* (S. 49–69). Cambridge: Cambridge University Press.

Schnotz, W. (2006). Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik. In H. Blühdorn, E. Breindl & U. H. Waßner (Hrsg.), *Text - Verstehen. Grammatik und darüber hinaus* (S. 222–238). Berlin: Walter de Gruyter.

Schnotz, W. & Dutke, S. (2004). Kognitionspsychologische Grundlagen der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz* (S. 61–99). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schnotz, W., Horz, H. McElvany, N., Schroeder, S., Ullrich, M., Baumert, J., Hachfeld, A. & Richter, T. (2010). Das BITE-Projekt. Integrative Verarbeitung von Bildern und Texten in der Sekundarstufe I. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansat-*

zes. Zeitschrift für Pädagogik, 56. Beiheft (S. 143–153). Weinheim: Beltz.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Hrsg.) (2005a). *Beschlüsse der Kultusministerkonferenz – Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10).* München: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Hrsg.) (2005b). *Beschlüsse der Kultusministerkonferenz - Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss (Jahrgangsstufe 10).* München: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Hrsg.) (2005c). *Beschlüsse der Kultusministerkonferenz - Bildungsstandards im Fach Physik für den Mittleren Schulabschluss (Jahrgangsstufe 10).* München: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland in Zusammenarbeit mit dem Institut zur Qualitätsentwicklung im Bildungswesen (Hrsg.) (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* München: Luchterhand.

Simon, H. A. (1978). On the forms of mental representation. In C. W. Savage (Hrsg.), *Minnesota studies in the philosophy of science. Vol. IX: Perception and cognition. Issues in the foundations of psychology* (S. 3–18). Minneapolis: University of Minnesota Press.

Stanat, P. & Schneider, W. (2004). Schwache Leser unter 15-jährigen Schülerinnen und Schülern in Deutschland: Beschreibung einer Risikogruppe. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz: Vertiefende Analysen im Rahmen von PISA-2000* (S. 243–273). Wiesbaden: Verlag für Sozialwissenschaften.

Stecher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*, 1–14.

Sumfleth E. & Schüttler, S. (1995). Linguistische Textverständlichkeitskriterien – Helfen sie bei der Darstellung chemischer Inhalte weiter? *Zeitschrift für Didaktik der Naturwissenschaften, 1*, 55–72.

Trautwein, U., Köller, O., Lehmann, R. & Lüdke, O. (2007). Öffnung von Bildungswegen, erreichtes Leistungsniveau und Vergleichbarkeit von Abschlüssen. In U. Trautwein, O. Köller, R. Lehmann & O. Lüdke (Hrsg.), *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten* (S.11–29). Münster: Waxmann.

Upmeier zu Belzen, A. & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41–57.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Walpuski, M., Kampa, N., Kauertz, A. & Wellnitz, N. (2008). Evaluation der Bildungsstandards in den Naturwissenschaften. *Der mathematische und naturwissenschaftliche Unterricht, 61*, 323–326.

Walpuski, M., Kauertz, A., Kampa, N., Fischer, H. E., Mayer, J., Sumfleth, E. & Wellnitz, N. (2010). ESNaS – Evaluation der Standards für die Naturwissenschaften in der Sekundarstufe I. In A. Gehrmann, U. Hericks, M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle – Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 171–184). Bad Heilbrunn: Klinkhardt.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung.* Berlin: Logos.

Wellnitz, N. & Mayer, J. (2008). Evaluation von Kompetenzstruktur und -niveaus zum Beobachten, Vergleichen, Ordnen und Experimentieren. In D. Krüger, A. Upmeier zu Belzen, T. Riemeier & K. Niebert (Hrsg.), *Erkenntnisweg Biologiedidaktik 7* (S. 129–143). Hannover: Universitätsdruckerei Kassel.

Wellnitz, N. & Mayer, J. (2011, September). *Erfassung und Förderung wissenschaftsmethodischer Kompetenzen.* Paper presented at the meeting of the Verband Biologie, Biowissenschaften & Biomedizin in Deutschland (VBIO), Bayreuth, Germany.

Wittwer, J., Saß, S. & Prenzel, M. (2008). Naturwissenschaftliche Kompetenz im internationalen Vergleich: Testkonzeption und Ergebnisse. In

W. Bos, M. Bonsen, J. Baumert, M. Prenzel, C. Selter & G. Walther (Hrsg.), *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 87–124). Münster: Waxmann.

Anhang A

Testmaterialien

Die folgenden Seiten enthalten Instruktionen für die Testdurchführung sowie Auszüge aus den Testheften zur Erfassung naturwissenschaftlicher Kompetenz in Biologie. Die Testaufgaben selbst sind Bestandteil der Evaluation der Bildungsstandards in den Naturwissenschaften, unterliegen als solche der Geheimhaltung und sind aus diesem Grund nicht abgedruckt. Hiervon ausgenommen ist lediglich die Aufgabe "Wasserflöhe", die in leicht überarbeiteter Form in der ESNaS-Pilotierungsstudie zum Einsatz gekommen ist und vom IQB zur Veröffentlichung freigegeben wurde. Neben den ESNaS-Aufgaben enthielten die Testhefte dieser Arbeit die Aufgabe "Obstkorb", die zwar nach den ESNaS-Konstruktionsprinzipien erstellt wurde, jedoch nicht Bestandteil der Evaluation der Bildungsstandards war (vgl. Abbildung 4.1 in dieser Arbeit). Der Anhang umfasst:

1. Instruktionen für die Testleiterinnen und Testleiter der Vorstudien
2. Deckblatt der Testhefte zur Erfassung naturwissenschaftlicher Kompetenz im Fach Biologie
3. Personenfragebogen der Vorstudien
4. Personenfragebogen der Hauptstudie
5. Beispielaufgabe "Wasserflöhe" mit je drei Items in den Versionen *Text mit Abbildungen* und *Text ohne Abbildungen*

Hinweise für die Testleiterin / den Testleiter

Vielen Dank, dass sich Ihre Klasse an unserer Studie beteiligt! Die vorliegenden Aufgaben dienen der Entwicklung eines Tests, der auch bei Schülerinnen und Schülern mit geringer Lesekompetenz eine zuverlässige und faire Erfassung naturwissenschaftlicher Kompetenz ermöglichen soll. Es können alle Schülerinnen und Schüler der neunten Klasse teilnehmen.

Bevor es losgeht, füllen bitte alle beteiligten Lehrkräfte die beiliegende Vertraulichkeitsvereinbarung aus und senden diese nach dem Test zusammen mit den Testheften an uns zurück (frankierte Rückumschläge liegen bei).

Vor Beginn der Schulstunde wird pro Schülerin bzw. Schüler ein **„Test zum wissenschaftlichen Arbeiten - Biologie“** ausgeteilt. Es gibt zwei unterschiedliche Testhefte: **mit** und **ohne** Abbildungen in den Aufgaben. Diese sind bereits gemischt und können in der Reihenfolge ausgeteilt werden, in der sie im Paket liegen.

Die Unterrichtsstunde sollte dann wie folgt ablaufen:

①	Beginn der Unterrichtsstunde: Das Deckblatt auf dem **„Test zum wissenschaftlichen Arbeiten - Biologie“** wird von der Lehrkraft laut vorgelesen. Anschließend wird zum Umblättern aufgefordert.
②	Die Schülerinnen und Schüler werden gebeten, ihre Daten einzutragen. Das meiste erklärt sich von alleine. Nur beim letzten Punkt (dem persönlichen „Code“) sollte nachgefragt werden, ob alle verstanden haben, was sie dort eintragen sollen.
③	Es wird zum Umblättern und zum Anfangen aufgefordert. Der Test dauert 25 Minuten und soll unbedingt pünktlich beendet werden (auch, wenn einzelne Schülerinnen und Schüler noch nicht fertig sind). **Die Testhefte werden eingesammelt!**
④	Pro Schülerin / Schüler wird nun ein Testheft **„LGVT 6-12“** ausgeteilt.
⑤	Auf dem Deckblatt ist **genau der selbe Code wie im ersten Testheft** einzutragen - siehe Punkt ②. Die anderen Felder (Geburtsdatum usw.) bleiben leer. Der Code dient dem anonymen Vergleich beider Tests.
⑥	Das Deckblatt des **„LGVT 6-12“** wird von der Lehrkraft laut vorgelesen. Das Beispiel wird gemeinsam bearbeitet. Anschließend wird zum Umblättern und Anfangen aufgefordert.
⑦	Der **„LGVT 6-12“** dauert 4 Minuten und soll unbedingt pünktlich beendet werden. Der Test ist so angelegt, dass man in dieser Zeit unmöglich fertig werden kann.
⑧	Alle Testhefte werden eingesammelt und in dem beiliegenden Rückumschlag an uns zurückgesendet. Wichtig ist, dass in beiden Heften eines Schülers / einer Schülerin jeweils genau derselbe Code eingetragen ist.

Experimentieren in Biologie

Im folgenden Test werden einfache biologische Experimente beschrieben. Deine Aufgabe ist es, verschiedene wissenschaftliche Fragestellungen zu diesen Experimenten zu beantworten.

Hier ein einfaches Beispiel:

Peter pflanzt zwei Tomatenpflanzen in zwei gleich große Blumentöpfe mit Erde. Die eine Pflanze gießt er regelmäßig mit Wasser, die andere nicht. Er beobachtet über einen längeren Zeitraum, wie sich die beiden Pflanzen entwickeln.

Welche Hypothese (Vermutung) kann Peter mit diesem Experiment prüfen?

Kreuze an.

- ☐ Pflanzen brauchen *Licht* zum Überleben.
- ☐ Pflanzen brauchen *Kohlenstoffdioxid* zum Überleben.
- ☐ Pflanzen brauchen *Wasser* zum Überleben.
- ☐ Pflanzen brauchen *Dünger* zum Überleben.

Der Unterschied zwischen den beiden Pflanzen ist, dass die eine mit Wasser gegossen wird, die andere nicht. Mit seinem Experiment kann Peter also prüfen, ob Pflanzen Wasser zum Überleben brauchen. Antwort „C“ ist deshalb richtig. Peters Hypothese lautet:

☒ Pflanzen brauchen *Wasser* zum Überleben.

Die Aufgaben im Test sind etwas schwerer als das Beispiel. Versuche, so viele wie möglich zu lösen. Es ist immer nur eine Antwort richtig. Du hast 25 Minuten Zeit.

Bitte blättere erst um, wenn Du dazu aufgefordert wirst!

...einen Moment noch!

Bitte trage erst die folgenden Informationen ein, bevor es losgeht:

Wie alt bist Du?

☐ 10 ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17

In welche Klasse gehst Du?

☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

Welche Sprache wird bei Euch zu Hause am meisten gesprochen?

Geschlecht:

☐ weiblich ☐ männlich

Schulform:

☐ Hauptschule ☐ Realschule ☐ Gymnasium

Die folgenden 6 Buchstaben und Zahlen sind Dein persönlicher Code. Dieser muss nachher auch noch auf einem anderen Fragebogen eingetragen werden!

Der Vorname Deiner Mutter beginnt mit den zwei Buchstaben	*Der Vorname Deines Vaters beginnt mit den zwei Buchstaben*	*Dein Geburtstag ist der…*
☐☐	☐☐	☐☐
z.B. „Anna“ → A N	z.B. „Peter“ → P E	z.B. 14. Januar → 1 4

Warte, bis Deine Lehrerin / Dein Lehrer Dich auffordert, umzublättern. Dann kannst Du beginnen, die Aufgaben zu bearbeiten. Arbeite zügig und überspringe keine Aufgabe.

Versuche auch, die Fragen auf der rechten Seite immer zügig zu beantworten und Dich nicht zu lange damit aufzuhalten.

Viel Erfolg!

Einen Moment noch!

Bitte trage erst die folgenden Informationen ein, bevor es losgeht:

Wie alt bist Du?

☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17

In welche Klasse gehst Du?

☐ 7. ☐ 8. ☐ 9. ☐ 10.

Welche Sprache wird bei Euch zu Hause am meisten gesprochen?

.....

Wird bei Euch zu Hause noch eine zweite Sprache gesprochen?

☐ nein ☐ ja, und zwar: .....

Geschlecht:

☐ weiblich ☐ männlich

Schulform:

☐ Hauptschule ☐ Realschule ☐ Gymnasium

letzte Zeugnisnote im Fach …

Biologie: ..... Deutsch: ..... Mathematik: .....

Hast Du schon einmal ein Schuljahr wiederholen müssen? (Angabe freiwillig)

☐ nein ☐ ja, und zwar wegen einer schlechten Note im Fach: .....

Die folgenden 6 Buchstaben und Zahlen sind Dein persönlicher Code. Dieser muss nachher auch noch auf einem anderen Fragebogen eingetragen werden!

Der Vorname Deiner Mutter beginnt mit den zwei Buchstaben	*Der Vorname Deines Vaters beginnt mit den zwei Buchstaben*	*Dein Geburtstag ist der…*
☐☐	☐☐	☐☐
z.B. „Anna“ → A N	z.B. „Peter“ → P E	z.B. 14. Januar → 1 4

Warte, bis Deine Lehrerin / Dein Lehrer Dich auffordert, umzublättern. Dann kannst Du beginnen, die Aufgaben zu bearbeiten. Arbeite zügig und überspringe keine Aufgabe.

Viel Erfolg!

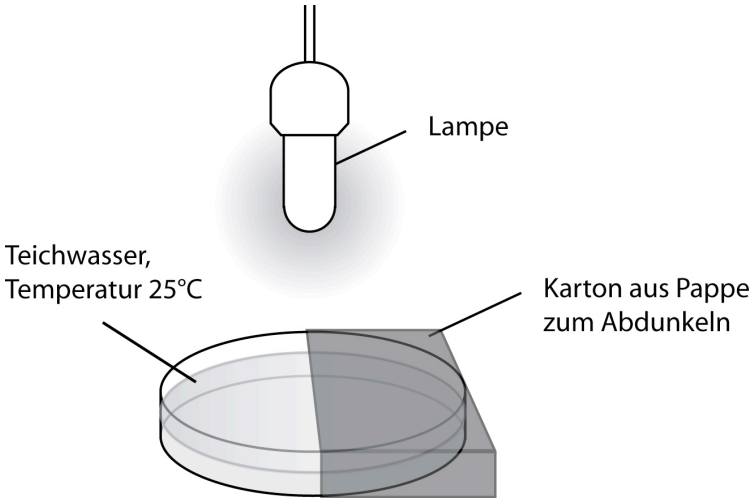
Bitte blättere erst um, wenn Du dazu aufgefordert wirst!

Wasserflöhe

Einige Fische ernähren sich von Wasserflöhen. Diese Kleinkrebse kann man an unterschiedlichen Stellen in einem Teich antreffen.

Christoph hat schon oft Wasserflöhe in einem Teich beobachtet. Er hat festgestellt, dass sich Wasserflöhe häufig an hellen, warmen Stellen aufhalten. Man findet sie oft im flachen Wasser in der Nähe von Wasserpflanzen.

Um seine Beobachtung wissenschaftlich zu überprüfen, führt Christoph folgenden Versuch durch:



In das Teichwasser in der Schale gibt er zehn Wasserflöhe.

Welcher Fragestellung will Christoph mit diesem Versuch nachgehen?

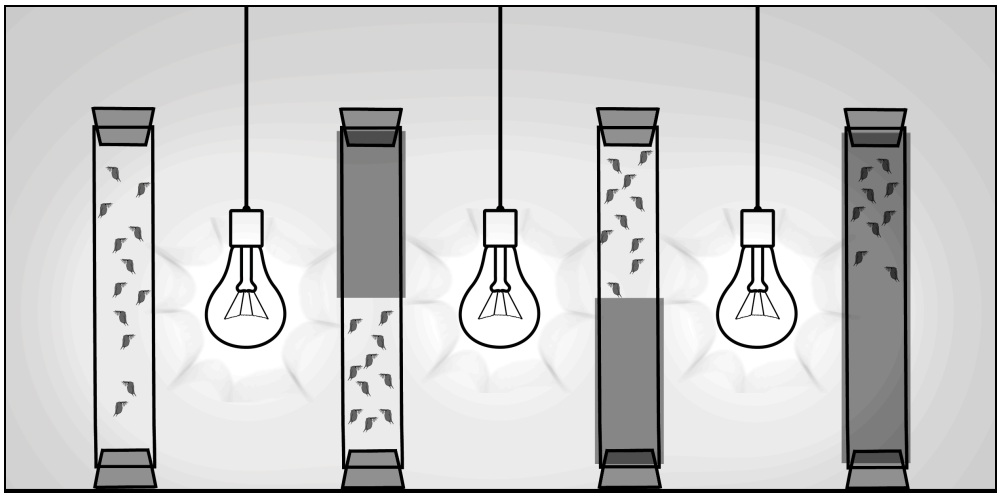
Kreuze an.

- ☐ Bevorzugen Wasserflöhe helle oder dunkle Stellen?
- ☐ Halten sich Wasserflöhe bevorzugt in der Nähe von Wasserpflanzen auf?
- ☐ Findet man Wasserflöhe meist im flachen Wasser?
- ☐ Bevorzugen Wasserflöhe warmes oder kaltes Wasser?

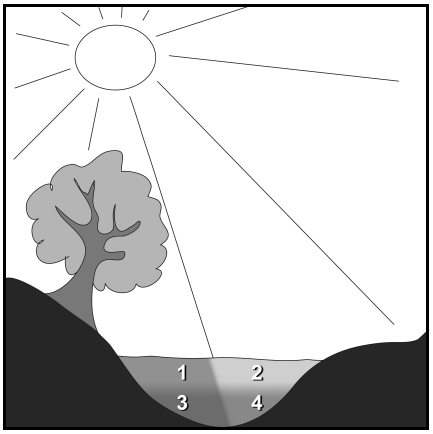
Wasserflöhe (Fortsetzung)

Christoph vermutet, dass noch weitere Bedingungen für den Aufenthaltsort von Wasserflöhen wichtig sind. Deshalb führt er einen weiteren Versuch durch.

Er füllt vier Glasröhren mit Wasser und setzt Wasserflöhe hinein. Die Röhren deckt er teilweise oder ganz mit lichtundurchlässigem Papier ab. Dann beobachtet Christoph, wo sich die Wasserflöhe aufhalten:



Christoph will die Ergebnisse dieses Experiments auf einen echten Teich übertragen:



Welche Schlussfolgerung kann er aus seinem Experiment ableiten?

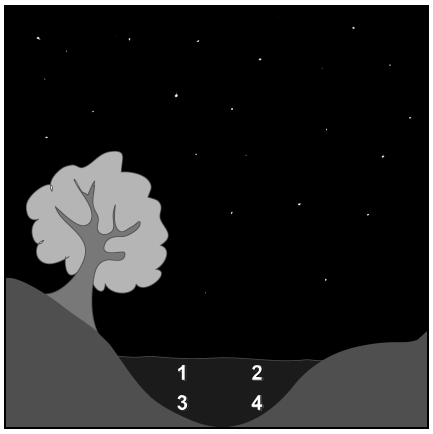
Kreuze an.

Tagsüber müssten sich die meisten Wasserflöhe ...

- ☐ ... in Bereich 1 aufhalten.
- ☐ ... in Bereich 2 aufhalten.
- ☐ ... in Bereich 3 aufhalten.
- ☐ ... in Bereich 4 aufhalten.

Wasserflöhe (Fortsetzung)

Christoph will die Ergebnisse des Experiments (siehe letzte Seite) auf einen echten Teich übertragen.



Welche Schlussfolgerung kann er aus seinem Experiment ableiten?

Kreuze an.

In einer dunklen Nacht müssten sich die meisten Wasserflöhe ...

- ☐ ... in Bereich 1 und 2 aufhalten.
- ☐ ... in Bereich 3 und 4 aufhalten.
- ☐ ... in der Mitte des Teiches aufhalten
- ☐ ... in allen Bereichen des Teiches aufhalten.

Wasserflöhe

Einige Fische ernähren sich von Wasserflöhen. Diese Kleinkrebse kann man an unterschiedlichen Stellen in einem Teich antreffen.

Christoph hat schon oft Wasserflöhe in einem Teich beobachtet. Er hat festgestellt, dass sich Wasserflöhe häufig an hellen, warmen Stellen aufhalten. Man findet sie oft im flachen Wasser in der Nähe von Wasserpflanzen.

Um seine Beobachtung wissenschaftlich zu überprüfen, führt Christoph folgenden Versuch durch:

Er gibt 25°C warmes Teichwasser in eine flache Schale aus Glas. Die Hälfte der Schale bedeckt er mit einem dunklen Karton aus Pappe. Über die Schale stellt er eine hell leuchtende Lampe.

In das Teichwasser in der Schale gibt er zehn Wasserflöhe.

Welcher Fragestellung will Christoph mit diesem Versuch nachgehen?

Kreuze an.

- ☐ Bevorzugen Wasserflöhe helle oder dunkle Stellen?
- ☐ Halten sich Wasserflöhe bevorzugt in der Nähe von Wasserpflanzen auf?
- ☐ Findet man Wasserflöhe meist im flachen Wasser?
- ☐ Bevorzugen Wasserflöhe warmes oder kaltes Wasser?

Wasserflöhe (Fortsetzung)

Christoph vermutet, dass noch weitere Bedingungen für den Aufenthaltsort von Wasserflöhen wichtig sind. Deshalb führt er einen weiteren Versuch durch.

Er füllt vier Glasröhren mit Wasser und setzt Wasserflöhe hinein. Die Röhren deckt er teilweise oder ganz mit lichtundurchlässigem Papier ab. Dann beobachtet Christoph, wo sich die Wasserflöhe aufhalten:

Röhre 1 ist nicht mit Papier abgedeckt. In der Röhre ist es überall hell. Die Wasserflöhe halten sich überall in der Röhre auf.

Röhre 2 ist oben mit Papier abgedeckt. In der Röhre ist es nur unten hell. Die Wasserflöhe halten sich nur unten in der Röhre auf.

Röhre 3 ist unten mit Papier abgedeckt. In der Röhre ist es nur oben hell. Die Wasserflöhe halten sich nur oben in der Röhre auf.

Röhre 4 ist vollständig mit Papier abgedeckt. In der Röhre ist es überall dunkel. Die Wasserflöhe halten sich nur oben in der Röhre auf.

Christoph will die Ergebnisse dieses Experiments auf einen echten Teich übertragen.

Welche Schlussfolgerung kann er aus seinem Experiment ableiten?

Kreuze an.

Tagsüber müssten sich die meisten Wasserflöhe ...

- ☐ ... an schattigen Stellen nahe der Wasseroberfläche aufhalten.
- ☐ ... an sonnigen Stellen nahe der Wasseroberfläche aufhalten.
- ☐ ... an schattigen Stellen in der Nähe des Teichgrundes aufhalten.
- ☐ ... an sonnigen Stellen in der Nähe des Teichgrundes aufhalten.

Wasserflöhe (Fortsetzung)

Christoph will die Ergebnisse des Experiments (siehe letzte Seite) auf einen echten Teich übertragen.

Welche Schlussfolgerung kann er aus seinem Experiment ableiten?

Kreuze an.

In einer dunklen Nacht müssten sich die meisten Wasserflöhe ...

- ☐ ... nahe unter der Wasseroberfläche aufhalten.
- ☐ ... in der Nähe des Teichgrundes aufhalten.
- ☐ ... in der Mitte des Teiches aufhalten.
- ☐ ... in allen Bereichen des Teiches aufhalten.

Abbildungsverzeichnis

Tabellenverzeichnis