# Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests

Dissertation zur Erlangung
des Doktorgrads der Philosophischen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Regine Asseburg

Kiel

(2011)

Erstgutachter: Prof. Dr. Jens Möller

Zweitgutachter: Prof. Dr. Andreas Frey

Tag der mündlichen Prüfung: 15.06.2011

Durch den zweiten Prodekan, Prof. Dr. Michael Düring,
zum Druck genehmigt am: 16.06.2011

# Inhaltsverzeichnis

# 1 Einleitung

Der Einsatz von Leistungstests gehört in einer leistungsorientierten Gesellschaft wie Deutschland, ebenso wie in anderen hoch entwickelten Staaten, zum Alltag. Leistungsdiagnostik wird in den unterschiedlichsten Bereichen des Lebens durchgeführt. So müssen sich Kinder und Jugendliche im Verlauf der Schulzeit einer Vielzahl von Leistungstests unterziehen, sei es in Form von Klassenarbeiten und Klausuren oder in Form von sportbezogenen Leistungstests wie den Bundesjugendspielen. Einige Personen nehmen darüber hinaus freiwillig an weiteren Leistungstests, beispielsweise an Wettbewerben wie Jugend musiziert, teil. Die Leistungsorientierung unserer modernen Wissensgesellschaft setzt sich im Erwachsenenalter fort. Zum Beispiel beruhen viele Auswahlverfahren für Ausbildungs- oder Studienplätze auf Leistungstests unterschiedlichster Art. Auch während der Ausbildung müssen in der Regel mehrfach Leistungstests bestanden werden. So selbstverständlich heutzutage lebenslanges Lernen erwartet wird, so selbstverständlich gehören Leistungsprüfungen zur Quantifizierung des Gelernten im Verlauf des Lebens zu unserem Alltag.

Der Einsatz von Leistungstests ist jedoch nicht nur im praktischen Leben, sondern auch in empirisch arbeitenden Wissenschaftsdisziplinen wie der empirischen Bildungsforschung üblich. Seit Deutschland Ende der 1990er Jahre erstmals an einer groß angelegten internationalen Schulleistungsstudie, der *Trends in International Mathematics and Science Study* (TIMSS), teilnahm, hat die Verbreitung groß angelegter Vergleichsstudien in Deutschland stark zugenommen (Stanat, 2008). So führte die Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) nach dem so genannten PISA-Schock im Jahr 2001 (*Programme for International Student Assessment*; OECD, 1999) flächendeckende Bildungsstandards in verschiedenen Schulfächern und für verschiedene Schulabschlüsse und -übergänge ein, um die Qualität des deutschen Schulsystems kontinuierlich zu prüfen und zu entwickeln (Köller, 2010). Auch die vor wenigen Jahren in Deutschland ins Leben gerufene *National Educational Panel Study* (NEPS), mit der Bildungsprozesse über die gesamte Lebensspanne hinweg untersucht werden, zeugt von der immensen Bedeutung, die die quantitative Leistungsdiagnostik im Bereich der Bildungsforschung hat.

Die Durchführung groß angelegter Vergleichsstudien bringt einen hohen administrativen, personellen und finanziellen Aufwand mit sich. Angesichts knapper öffentlicher Ressourcen und der hohen Belastung der Teilnehmerinnen und Teilnehmer durch die Testungen stellt sich die Frage, wie solche Studien möglichst effizient durchgeführt werden können. Aus psychometrischer Sicht bietet das so genannte computerisierte adaptive Testen (CAT) eine Möglichkeit, groß angelegte Studien sehr ökonomisch durchzuführen. Im CAT werden die Aufgaben, die einer Testperson vorgegeben werden, individuell auf Basis des gezeigten Antwortverhaltens ausgewählt, so dass niemand viel zu schwierige oder viel zu einfache Aufgaben beantworten muss. Dieser Grundgedanke des CAT erscheint intuitiv einleuchtend, und im Prinzip wird in den meisten mündlichen Prüfungen nach genau diesem Schema vorgegangen. Denn Aufgaben oder Fragen, deren Schwierigkeiten weit entfernt sind von dem Leistungsniveau der zu prüfenden Person, liefern kaum Informationen über die Höhe der zu beurteilenden Leistungsausprägung und führen zu einer unnötigen Verlängerung der Prüfung. CAT stellt durch die individuell angepasste Aufgabenauswahl eine erhebliche zeitliche Verkürzung der Testung und damit eine Verminderung des mit der Testung verbundenen Aufwands

bei gleich bleibender Messpräzision in Aussicht. Die Messeffizienzsteigerung durch CAT im Vergleich zu herkömmlichem Testen konnte in vielen Studien nachgewiesen werden (vgl. Frey, 2007).

Die psychometrischen Vorzüge von CAT sind mittlerweile gründlich beforscht. Ungeklärt ist allerdings, wie sich CAT bei den Personen, die den Test bearbeiten, auf die Motivation zur Testbearbeitung auswirkt. Auch existiert bisher kein Modell, das die in einer Leistungstestsituation aktuell ablaufenden motivationalen Prozesse hinreichend genau abbildet. Die Untersuchung der Auswirkungen von CAT auf die Motivation zur Testbearbeitung erscheint wichtig, da die gezeigte Leistung in einem Test nicht nur von kognitiven Merkmalen abhängt, sondern auch mit der individuellen, aktuellen Motivation zur Testbearbeitung im Zusammenhang steht. Das bedeutet, dass Testergebnisse nicht zwingend die maximale Leistung einer Person widerspiegeln. Dennoch wird genau dies in der Regel implizit unterstellt. Die Testergebnisse werden als maximale Leistung interpretiert, und mögliche Konsequenzen aus den Testergebnissen, beispielsweise die Zu- oder Absage eines Studienplatzes, werden unter dieser impliziten Annahme gezogen. Entspricht die gezeigte Leistung nicht der maximalen Leistung und wird aber als solche interpretiert, muss die Validität der aus den Ergebnissen gezogenen Schlussfolgerungen in Frage gestellt werden. Falls bestimmte Personengruppen außerdem durch bestimmte Testmerkmale wie die Adaptivität eines Tests systematisch demotiviert werden, kann dies eine Beeinträchtigung der Testfairness bedeuten. Angesichts der viel versprechenden psychometrischen und ökonomischen Vorzüge von CAT auf der einen Seite und den ungeklärten Fragen zu dessen motivationalen Auswirkungen auf der anderen Seite erscheint es wichtig und notwendig, die motivationalen Prozesse in adaptiven und nicht-adaptiven Leistungstests näher zu beleuchten.

Die vorliegende Arbeit entsteht demzufolge aus zwei Beweggründen heraus: Erstens soll ein Modell zu Motivationsprozessen im Testverlauf abgeleitet werden, um die bestehende Lücke in der Motivationsforschung zu schließen (theoretische Relevanz der Arbeit; vgl. auch Brunstein & Heckhausen, 2006; Cole, Bergin & Whittaker, 2008). Zweitens sollen die Effekte verschiedener Test-, Situations- und Personenmerkmale auf die Motivation zur Testbearbeitung systematisch untersucht werden. Das Ziel ist, eine wissenschaftlich fundierte Empfehlung für die Konzeption groß angelegter Studien auszusprechen, die eine möglichst hohe Messeffizienz und gleichzeitig eine hohe Motivation der Testpersonen gewährleistet (praktische Relevanz der Arbeit). Denn: „It may be time to turn the focus to the person as well as the score" (Wolf & Smith, 1995, S. 240).

Der erste Teil der Arbeit stellt die theoretischen Grundlagen bereit, die für ein Verständnis der Fragestellungen und Hypothesen notwendig sind. Er beginnt mit Erläuterungen zu CAT und einer kritischen Auseinandersetzung mit CAT (Abschnitt 2). Anschließend werden die Begriffe Leistungsmotivation und Motivation zur Testbearbeitung (als spezifische Form der Leistungsmotivation) erörtert, und es werden Erwartung-Wert-Modelle der Leistungsmotivation vorgestellt (Abschnitt 3). Diese Modelle haben sich insbesondere im schulischen Bereich zur Vorhersage der Leistungsmotivation und der Leistung etabliert. Vor dem Hintergrund dieser theoretischen Auseinandersetzung mit Leistungsmotivation folgt eine kritische Analyse des empirischen Forschungsstands zur Motivation zur Testbearbeitung (Abschnitt 4). Auf Basis der theoretischen Konzeptionen und der empirischen Befunde werden die Fragestellungen und die Hypothesen der Arbeit abgeleitet (Abschnitt 5).

Es schließt sich der empirische Teil der Arbeit an, welcher der Prüfung der Hypothesen dient. Zunächst werden detaillierte Informationen zur Datenerhebung und -analyse gegeben (Abschnitt 6). Es folgt die Darstellung der Ergebnisse zu den Hypothesen (Abschnitt 7). Abschließend werden die Ergebnisse unter Berücksichtigung der theoretischen Annahmen und des bisherigen empirischen Forschungsstands kritisch diskutiert (Abschnitt 8). Es wird ein Resümee der Arbeit gezogen und ein Ausblick auf mögliche zukünftige, sich anknüpfende Forschungsfragen gegeben (Abschnitt 9).

# 2 Computerisiertes adaptives Testen

Die vorliegende Arbeit beschäftigt sich mit Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests. Adaptives Testen beruht auf einem komplexen, mathematisch anspruchsvollen Testalgorithmus und wird daher in der Regel computerisiert, also ausschließlich am Computer, durchgeführt. Deswegen beschränkt sich die Arbeit auf computerisierte Tests. *Computerisiertes adaptives Testen* (CAT) gilt zunehmend als attraktive Art zu testen (Georgiadou, Triantafillou & Economides, 2006; Lunz, Bergstrom & Gershon, 1994; Zenisky & Sireci, 2002). Es kann jedoch im deutschen Sprachraum noch nicht als gängiges Testverfahren gelten. Daher werden in diesem ersten Abschnitt die Grundlagen von CAT vorgestellt. Das Verständnis von CAT ist wichtig, um die psychometrischen Vorteile dieses Testalgorithmus zu erkennen und um mögliche Effekte von CAT auf die Motivation zur Testbearbeitung nachzuvollziehen (vgl. Hypothesen, Abschnitt 5).

In einem herkömmlichen Leistungstest wird allen Testpersonen dieselbe, vorab festgelegte Menge an Aufgaben in derselben, vorab festgelegten Reihenfolge vorgegeben (sogenanntes *fixed item testing*, FIT). Im CAT hingegen werden die Aufgaben individuell auf Basis des gezeigten Antwortverhaltens einer Testperson während der Testbearbeitung sukzessive ausgewählt und vorgegeben. Die Idee des adaptiven Testens ist dabei so alt wie das Testen in der psychologischen Diagnostik an sich. Bereits Binet und Simon (1904) gaben in ihrem Intelligenztest nicht jeder Testperson dieselben Aufgaben vor, sondern wählten die nach Schwierigkeit sortierten Aufgaben anhand eines vorgegebenen Schemas auf Basis des Antwortverhaltens der Testpersonen individuell aus. Hinter diesem Vorgehen steckt der Gedanke, dass viel zu schwierige oder viel zu einfache Aufgaben kaum Aussagen darüber erlauben, wie hoch die Leistungsausprägung einer Testperson ist. Kann eine Testperson beispielsweise alle vorgegebenen Aufgaben in einem bestimmten Test korrekt beantworten, liegt zwar eine gewisse Information über die Leistungsausprägung dieser Person vor, doch verbleibt die maximal mögliche Leistung dieser Person unbekannt. Gerade diese maximale Leistung soll jedoch in der Regel mit einem Leistungstest erfasst werden (Cronbach, 1970; Wise & DeMars, 2005). Daher ist ein Test, der die Testperson dauerhaft unter- oder überfordert, nicht optimal. Zudem kann die wiederholte Vorgabe von Aufgaben mit inadäquater Schwierigkeit bei der Testperson Ermüdung oder Frustration hervorrufen (Lilley, Barker & Britton, 2004). Dies kann der Bereitschaft, maximale Leistung zu zeigen, zuwiderlaufen (Csikszentmihalyi, 1990; Linacre, 2000).

Adaptives Testen hat den Zweck, dass die Aufgaben, die eine Person bearbeitet, eine möglichst hohe Diagnostizität aufweisen. Dies bedeutet, dass die individuell ausgewählten Aufgaben möglichst viel Information über die zu messende Merkmalsausprägung dieser Person liefern (vgl. auch Abschnitt 2.2 zum Zusammenhang zwischen CAT und Messeffizienz, und Abschnitt 3.2.1.3 zur Bedeutsamkeit der Diagnostizität für die Motivation zur Testbearbeitung). Die Diagnostizität ist maximal, wenn die Schwierigkeit der Aufgabe der Merkmalsausprägung der Person entspricht (Wainer, 2000). Da in einem FIT in der Regel durchschnittlich nur wenige Aufgaben pro Person dieses Kriterium erfüllen, wird die Möglichkeit einer maximal informativen und effizienten Leistungsmessung nicht ausgeschöpft. Dafür wäre es notwendig, dass alle Personen der Stichprobe dieselbe Merkmalsausprägung aufweisen und diese der Schwierigkeit der vorgegebenen Aufgaben entspricht, was eine unrealistische Annahme ist.

Die Entstehung von CAT ist vor allem auf zwei Entwicklungen zurückzuführen: zum einen auf große Fortschritte im Bereich der statistischen Grundlagen seit den 1950er Jahren, die zur Formulierung der sogenannten Item-Response-Theorie geführt haben, auf der alle adaptiven Testverfahren beruhen (Lord, 1980; vgl. Abschnitt 2.1), und zum anderen auf die Entwicklung und rasante Verbreitung von Computern seit den 1980er Jahren, die bald darauf auch zur Testadministration eingesetzt wurden (vgl. van der Linden, 2008). Die Grundlagen der Item-Response-Theorie werden in Abschnitt 2.1 näher erläutert. Wie ein adaptiver Test entwickelt wird, was seine wesentlichen Bestimmungsstücke sind und weshalb adaptives Testen eine höhere Messeffizienz als herkömmliche, nicht-adaptive Testverfahren erreicht, wird in Abschnitt 2.2 dargelegt. In Abschnitt 2.3 werden Vor- und Nachteile dieser Art zu testen diskutiert, bevor in Abschnitt 2.4 verschiedene Anwendungen computerisierten adaptiven Testens in der empirischen Bildungsforschung vorgestellt werden.

## 2.1 Item-Response-Theorie

Aktuelle computerisierte adaptive Testverfahren beruhen auf der Item-Response-Theorie (IRT; z. B. van der Linden & Hambleton, 1997), deren Grundlagen im Folgenden dargestellt werden. Diese Grundlagen werden benötigt, um das Funktionsprinzip des adaptiven Testalgorithmus und die Zusammenhänge zwischen Merkmalsausprägung, Aufgabenschwierigkeit, Lösungswahrscheinlichkeit und diagnostischem Informationsgehalt einer Aufgabe zu verstehen. Die Grundüberlegungen zur IRT (auch als Probabilistische Testtheorie bekannt; Moosbrugger, 2002) entstanden in den 1950er Jahren. In den 1960er bis 1980er Jahren wurden diese Grundüberlegungen in ein formalisiertes Modell übersetzt und fortwährend statistisch elaboriert (van der Linden, 2005). Seit den 1990er Jahren erfährt die IRT, die als Ergänzung zur Klassischen Testtheorie (KTT) zu verstehen ist, weltweit Anwendung (Moosbrugger, 2007a; Rost, 1999; für weitergehende Informationen zur KTT siehe z. B. Moosbrugger, 2007b).

Die IRT geht davon aus, dass das Antwortverhalten einer Testperson nicht deterministisch bestimmt ist, sondern dass sich nur Wahrscheinlichkeitsaussagen über das Verhalten einer Person im Test treffen lassen. Die IRT versucht, die Verhaltensprognosen zu formalisieren und damit einer empirischen Überprüfung zugänglich zu machen. Sie beschreibt sozusagen mathematisch, was passiert, wenn eine Testperson auf eine Aufgabe trifft (Wainer & Mislevy, 2000; vgl. auch Weiss, 2004). Die IRT beruht auf strengen Annahmen: So müssen die Aufgaben eines Tests *homogen* sein, das heißt, alle Aufgaben müssen dieselbe latente Dimension messen (*Eindimensionalität*). Als latente Dimension versteht man dabei ein Merkmal, das sich der direkten Beobachtung entzieht und das mit Hilfe der beobachtbaren (manifesten) Antworten auf die Aufgaben eines Tests erfasst werden soll. Ein solches latentes Merkmal ist mathematische Kompetenz. Eine hinreichende Bedingung für das Vorliegen von Aufgabenhomogenität bezüglich einer latenten Dimension $\theta$ ist, dass sich Korrelationen zwischen den Aufgaben ausschließlich auf Unterschiede zwischen den Merkmalsausprägungen von Personen auf $\theta$ zurückführen lassen. Inhaltlich liegt der Bedingung der Aufgabenhomogenität die Annahme zugrunde, dass die Merkmalsausprägung das Antwortverhalten im Test bewirkt und dadurch die Korrelationen zwischen den Aufgaben hervorruft. Im Falle der mathematischen Kompetenz bedeutet dies beispielsweise, dass keine Korrelationen mehr zwischen den Aufgaben bestehen, wenn ausschließlich Personen mit gleich hoher mathematischer Kompetenz

an dem Test teilnehmen. Neben der zugrunde liegenden Dimension θ, die mit dem Test gemessen werden soll, teilen die Aufgaben keinerlei weitere Varianz; sie kovariieren darüber hinaus nicht. Dies bedeutet, dass die Antwort einer Testperson auf eine Aufgabe unabhängig von ihrer Antwort auf andere Aufgaben ist. Diese Annahme wird als *lokale stochastische Unabhängigkeit* bezeichnet. Sie ist die Grundvoraussetzung für Aufgaben, die in einem CAT verwendet werden (vgl. Frey, 2007). Liegt lokale stochastische Unabhängigkeit der Aufgaben vor, werden diese homogenen Testaufgaben als Indikatoren der latenten Dimension bezeichnet, und das manifeste Antwortverhalten auf diese Aufgaben wird als Indikator für die Ausprägung des latenten Personenmerkmals begriffen.

Die Darstellung in der vorliegenden Arbeit beschränkt sich auf das einfachste und geläufigste IRT-Modell, das *dichotome Latent-Trait-Modell* (vgl. Moosbrugger, 2002). *Dichotom* bedeutet, dass sich das Modell ausschließlich auf Aufgaben bezieht, für deren Beantwortung es nur zwei komplementäre, also sich einander ausschließende, Ereignisse gibt (z. B. richtig oder falsch). Aufgaben, für die es bei anteilig korrekter Beantwortung Teilpunkte gibt, werden durch dieses Modell nicht abgedeckt. Der Begriff *latent trait* bezieht sich auf die latente Dimension (englisch: trait = Merkmal). Das dichotome Latent-Trait-Modell geht davon aus, dass sich die Ausprägung einer Person *v* auf einer zu messenden, kontinuierlichen latenten Dimension θ durch einen einzigen Parameter (Personenparameter $\theta_v$) ausdrücken lässt. Diese Merkmalsausprägung $\theta_v$ bestimmt das Antwortverhalten der Person auf eine dichotome Aufgabe *i*. Je nach verwendetem IRT-Modell geschieht dies

- in Abhängigkeit von dem Aufgabenschwierigkeitsparameter[1] $b_i$ (1PL-Modell bzw. Rasch-Modell), der definiert ist als diejenige Merkmalsausprägung θ, bei der die Wahrscheinlichkeit, die Aufgabe korrekt zu beantworten, 50 Prozent beträgt;
- in Abhängigkeit von $b_i$ und dem Aufgabendiskriminationsparameter $a_i$ (2PL-Modell bzw. Birnbaum-Modell), der angibt, wie gut eine Aufgabe zwischen verschiedenen Merkmalsausprägungen der Personen trennt; oder
- in Abhängigkeit von $b_i$, $a_i$ und dem Rateparameter $c_i$ (3PL-Modell bzw. Rate-Modell von Birnbaum; vgl. Moosbrugger, 2007a), welcher die Wahrscheinlichkeit angibt, bei minimaler Merkmalsausprägung die Aufgabe dennoch durch Raten korrekt zu lösen.

Dem CAT in dieser Arbeit liegt das Rasch-Modell zugrunde (Rasch, 1960), daher beschränkt sich die weitere Darstellung auf dieses 1PL-Modell (mit $a_i = 1$, d. h. der Aufgabendiskriminationsparameter ist für alle Aufgaben konstant und auf eins festgelegt). Die Bezeichnung „1PL“ ist eine Abkürzung für *one parameter logistic* und leitet sich daraus ab, dass jede Aufgabe in diesem Modell durch nur einen Parameter, die Aufgabenschwierigkeit *b*, charakterisiert wird (Wainer & Mislevy, 2000). Die Wahrscheinlichkeit, dass eine Person *v* eine Testaufgabe *i* korrekt beantworten kann, ergibt sich allein aus der Differenz von $\theta_v$ und $b_i$.

[1] Nachfolgend wird statt „Aufgabenschwierigkeitsparameter“ der besseren Lesbarkeit halber der Begriff „Aufgabenschwierigkeit“ verwendet.

Das Rasch-Modell definiert für alle Aufgaben die gleiche logistische Funktion für die Wahrscheinlichkeit einer richtigen Antwort:

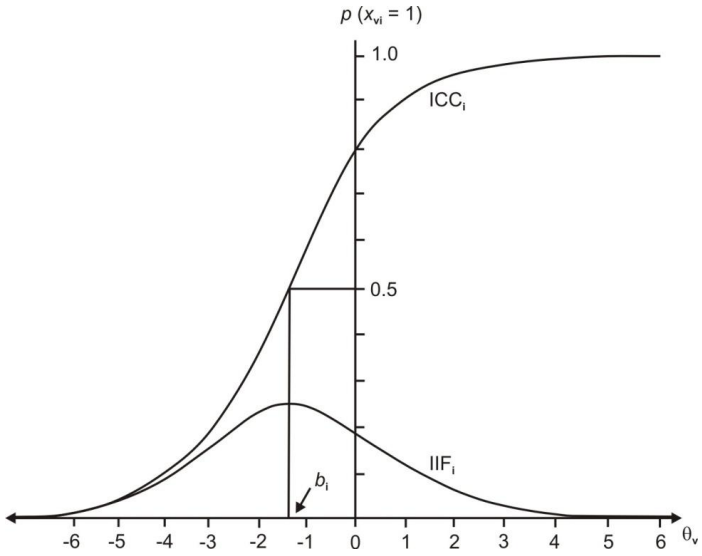
$$p\,(x_{\mathrm{vi}} = 1\,|\theta_{\mathrm{v}}, b_{\mathrm{i}}) = \frac{\mathrm{e}^{(\theta_{\mathrm{v}} - b_{\mathrm{i}})}}{1 + \mathrm{e}^{(\theta_{\mathrm{v}} - b_{\mathrm{i}})}} = \mathrm{logit}\,(\theta_{\mathrm{v}} - b_{\mathrm{i}})$$

Die Funktion beschreibt die Wahrscheinlichkeit *p*, dass die Antwort *x* einer Person *v* auf eine Aufgabe *i* richtig ist, bei gegebenem Personenparameter $\theta_v$ und gegebener Aufgabenschwierigkeit $b_i$. Wie man an der Gleichung sieht, ist die Differenz zwischen Personenparameter und Aufgabenschwierigkeit, $\theta_v$ - $b_i$, im Rasch-Modell das zentrale Bestimmungsstück zur Berechnung der Wahrscheinlichkeit, eine Aufgabe korrekt zu lösen. Da sich Personenparameter und Aufgabenschwierigkeit im Rahmen der IRT auf einer gemeinsamen Skala abbilden lassen (*joint scale*; van der Linden, 2008), lässt sich eine einfache Differenz zwischen diesen beiden Werten bilden. Die gemeinsame Skala wird in den logistischen IRT-Modellen auch als Logit-Skala bezeichnet (vgl. rechte Seite der Gleichung; z. B. Hartig & Höhler, 2009). Ein Logit ist dabei als Logarithmus des so genannten Wettquotienten von Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit definiert (Moosbrugger, 2007a; Rost, 2004):

$$\text{Logit: } \log \frac{p\,(x_{\mathrm{vi}} = 1)}{p\,(x_{\mathrm{vi}} = 0)}$$

Ist die Differenz zwischen Personenparameter und Aufgabenschwierigkeit positiv, übersteigt die individuelle Merkmalsausprägung die Aufgabenschwierigkeit, und die Wahrscheinlichkeit einer korrekten Antwort ist größer als 50 Prozent. Ist die Differenz negativ, übersteigt die Aufgabenschwierigkeit die individuelle Merkmalsausprägung, und die Wahrscheinlichkeit einer richtigen Antwort ist kleiner als 50 Prozent. Die IRT berücksichtigt folglich explizit die Möglichkeit, dass eine Testperson eine Aufgabe, die eigentlich „zu schwierig" für sie ist, dennoch lösen kann, und dass eine Testperson eine Aufgabe, die eigentlich „zu leicht" für sie ist, dennoch falsch beantworten kann. Je stärker $\theta_v$ und $b_i$ voneinander abweichen, umso unwahrscheinlicher ist allerdings bei negativer Differenz eine richtige Antwort und bei positiver Differenz eine falsche Antwort.

Die Wahrscheinlichkeitsfunktion des Rasch-Modells wird üblicherweise graphisch in Form einer itemcharakteristischen Funktion dargestellt (*Item-Characteristic Curve*, ICC; Abbildung 2.1). Auf der Abszisse wird dabei die Ausprägung der latenten Dimension und auf der Ordinate die Lösungswahrscheinlichkeit abgetragen. Die Aufgaben im Rasch-Modell haben alle die gleiche ICC, nur dass die Kurven entlang der Abszisse parallel verschoben sind. Man sagt auch, dass die Modellgleichung des Rasch-Modells eindeutig ist bis auf positiv lineare Transformationen (Moosbrugger, 2007a). Der parallele Verlauf der ICCs ermöglicht eine vorteilhafte Eigenschaft des Rasch-Modells und eine Grundvoraussetzung für Aufgaben, die zum computerisierten adaptiven Testen verwendet werden: die *spezifische Objektivität der Vergleiche*. Dies bedeutet, dass der Unterschied zwischen den Ausprägungen zweier Personen auf der latenten Dimension (z. B. Kompetenz) unabhängig von den spezifischen vorgegebenen Aufgaben festgestellt werden kann. Ebenso kann der Schwierigkeitsunterschied zwischen zwei Aufgaben unabhängig von den untersuchten Personen errechnet werden.




$\theta_v$: Merkmalsausprägung von Person *v*

$b_i$: Schwierigkeit von Aufgabe *i*

$p\ (x_{vi} = 1)$: Wahrscheinlichkeit für Person *v*, Aufgabe *i* richtig zu lösen

$ICC_i$: Itemcharakteristische Funktion von Aufgabe *i*

$IIF_i$: Iteminformationsfunktion von Aufgabe *i*

*Abbildung 2.1:* ICC und IIF einer Aufgabe im dichotomen Rasch-Modell.

Die Steigung einer ICC ist im Wendepunkt maximal und entspricht hier der Diskrimination der Aufgabe, die im Rasch-Modell für alle Aufgaben auf eins fixiert ist. Die Lokalisation des Wendepunkts markiert zudem auf der Abszisse die Aufgabenschwierigkeit sowie auf der Ordinate eine Lösungswahrscheinlichkeit von 50 Prozent für den Fall, dass die individuelle Merkmalsausprägung der Aufgabenschwierigkeit entspricht. Dass die Lösungswahrscheinlichkeit genau 50 Prozent beträgt, wenn die Differenz zwischen Personenparameter und Aufgabenschwierigkeit Null ist, lässt sich leicht aus der Formel des Rasch-Modells ableiten:

$$p\ (x_{\mathrm{vi}} = 1\ |\theta_{\mathrm{v}}, b_{\mathrm{i}}) = \frac{\mathrm{e}^{0}}{1 + \mathrm{e}^{0}} = \frac{1}{2}$$

An diesem Punkt liefert die Aufgabe die maximale Information über die Merkmalsausprägung der Person. Denn die Steigung der ICC lässt sich im Rasch-Modell in der so genannten Iteminformationsfunktion abbilden (IIF, vgl. Abbildung 2.1; Moosbrugger, 2002). Diese wird aus dem Produkt der beiden Wahrscheinlichkeiten errechnet, die Aufgabe richtig beziehungsweise falsch zu beantworten. Dieses Produkt, und damit ebenso die Iteminformation und die Steigung, ist maximal, wenn beide Faktoren maximal hoch sind. Dies ist der Fall, wenn beide den Wert 0.5 annehmen (für Details zur Berechnung von Iteminformationsfunktionen siehe z. B. Hambleton & Cook, 1977). Hier knüpft der Kerngedanke von CAT an: Üblicherweise werden die zu bearbeitenden Aufgaben in einem CAT individuell so ausgewählt, dass die Lösungswahrscheinlichkeit 50 Prozent beträgt. Auf diese Weise kann mit möglichst wenigen Aufgaben möglichst viel Information über die individuelle Merkmalsausprägung gewonnen werden (und damit eine hohe Messeffizienz erreicht werden; vgl. Abschnitt 2.3).

Auch die Personenparameter- und Aufgabenschwierigkeitsschätzung macht sich die lokale stochastische Unabhängigkeit der Aufgaben und damit die Gültigkeit des Multiplikationstheorems für unabhängige Ereignisse zunutze (vgl. hierzu Moosbrugger, 2007a). Unter Verwendung der so genannten Likelihood-Funktion

$$L = \prod_{v=1}^{N} \prod_{i=1}^{k} p\,(x_{\mathrm{vi}})$$

werden $\theta$ und $b$ über das wiederholte systematische Anwenden des Multiplikationstheorems auf Basis der vorliegenden Antworten so geschätzt, dass die Wahrscheinlichkeit der beobachteten Daten unter den Modellannahmen maximiert wird (so genannte *Maximum Likelihood*). Für die Schätzung der Personenparameter und der Aufgabenschwierigkeiten ist es bei gegebener lokaler stochastischer Unabhängigkeit unerheblich, welche Aufgaben von welchen Personen gelöst wurden. Allein die Anzahl der gelösten Aufgaben pro Person beziehungsweise die Anzahl der Personen mit korrekter Antwort pro Aufgabe ist ausreichend, um die Parameter zu schätzen. Die jeweiligen einfachen Summenwerte bilden daher so genannte *erschöpfende Statistiken* für die Parameterschätzung. Ob diese Eigenschaft, Antworten zu Summenwerten addieren zu können, gegeben ist und zu interpretierbaren Messwerten führt, ob also Modellkonformität besteht, kann im Rahmen des Rasch-Modells geprüft werden. Kann eine Person keine einzige Aufgabe oder aber alle Aufgaben korrekt beantworten, ist der Personenparameter über die Maximum-Likelihood-Methode nicht eindeutig bestimmbar, da er gegen -∞ beziehungsweise +∞ geht. In diesem Fall kann entweder auf die so genannte Gewichtete-Maximum-Likelihood-Methode oder auf so genannte Bayes-Schätzer zurückgegriffen werden (für Details siehe Rost, 2004, oder Segall, 2005).

Ein weiterer Vorteil des Rasch-Modells knüpft an die erschöpfenden Statistiken an: Die geschätzten Aufgabenparameter sind *stichprobenunabhängig*. Das heißt, die Parameterschätzung ist ohne Kenntnis der Personenparameter und ohne Annahmen über deren Verteilung möglich. Dies hat den Vorteil, dass individualisiert getestet werden kann. Die Stichprobenunabhängigkeit der Parameterschätzung ist ein Grund dafür, dass sich die Anwendung IRT-basierter Test- und Schätzverfahren in der empirischen Bildungsforschung etabliert hat. Sie ermöglicht den Einsatz so genannter Multi-Matrix-Designs, bei denen – ähnlich wie bei CAT – nicht alle Testpersonen dieselben Aufgaben bearbeiten, sondern jede Testperson lediglich ein Testheft mit einer Teilmenge der Aufgaben erhält. Die Zuteilung der Aufgaben auf die Testhefte erfolgt im Gegensatz zu CAT bereits vor Beginn des Tests auf Basis eines bestimmten Regelsystems (z. B. anhand von Youden-Squares; für einen Überblick zu Testheft-Designs siehe z. B. Frey, Hartig & Rupp, 2009). Multi-Matrix-Designs ermöglichen eine vergleichsweise ökonomische Erfassung von Schülerleistungen in groß angelegten Vergleichsstudien (vgl. PISA: Baumert, Artelt, Klieme & Stanat, 2001; TIMSS: Klieme, Baumert, Köller & Bos, 2000). Auch CAT macht sich die Stichprobenunabhängigkeit der Parameterschätzung und die übrigen vorteilhaften Eigenschaften rasch-homogener Tests zunutze. Somit ist für CAT eine IRT-konforme Testentwicklung eine notwendige Voraussetzung (vgl. Abschnitte 6.2.1 und 7.1).

## 2.2 Bestimmungsstücke computerisierten adaptiven Testens

Bei der Konstruktion eines CAT müssen einige Grundvoraussetzungen erfüllt sein. Die wesentlichen Bestimmungsstücke eines CAT lassen sich in Anlehnung an Thissen und Mislevy (2000) sowie Weiss und Kingsbury (1984) allgemein folgendermaßen skizzieren (vgl. Abschnitt 6.2.1 für Details zu dem CAT der vorliegenden Arbeit):

a) Item-Response-Modell:
   Ein CAT muss auf Basis der Item-Response-Theorie entwickelt werden (Hambleton, Swaminathan & Rogers, 1991). Im Rahmen der IRT muss das Messmodell spezifiziert werden, das dem CAT zugrunde gelegt wird (z. B. das Rasch-Modell). Die Entscheidung für das passende Modell sollte im Einzelfall anhand psychometrischer und inhaltlicher Kriterien getroffen werden.

b) Aufgabenpool:
   Um auf allen Bereichen der latenten Merkmalsdimension (z. B. auf allen Bereichen der Kompetenzverteilung) eine hinreichend präzise Parameterschätzung vornehmen zu können, muss ein Aufgabenpool vorliegen, der hinreichend viele Aufgaben mit Schwierigkeiten im gesamten Bereich möglicher Merkmalsausprägungen enthält. Diese Aufgaben müssen zudem alle das Kriterium der lokalen stochastischen Unabhängigkeit erfüllen (vgl. Abschnitt 2.1). Dies bedeutet, dass die Antwort auf eine Aufgabe unabhängig von der Antwort auf die anderen Aufgaben im Pool ist. Wenn eine Testperson beispielsweise die richtige Antwort zu Aufgabe A nicht weiß, darf ihr daraus kein Nachteil bei der Beantwortung von Aufgabe B erwachsen. Anders ausgedrückt darf Aufgabe B nicht auf Aufgabe A aufbauen. Berührt der Test verschiedene Inhaltsbereiche, sollten in jedem Inhaltsbereich ausreichend und ähnlich viele Aufgaben in dem gesamten Schwierigkeitsbereich zur Verfügung stehen, in dem Merkmalsausprägungen zu erwarten sind. Die Gesamtmenge an Aufgaben in einem CAT-Aufgabenpool sollte etwa acht- bis zwölfmal so groß sein wie die Anzahl der vorzugebenden Aufgaben (Faustregel nach de Ayala, 2009). Nach Weiss und Kingsbury (1984) kann ein Aufgabenpool mit 100 Aufgaben bereits zufriedenstellend sein, wenn die Aufgaben bezüglich ihrer Schwierigkeiten hinreichend gut verteilt sind.

c) Anfangsaufgabe (How to start?):
   Die Aufgabe, mit der begonnen wird, muss vorab festgelegt werden. In der Regel wird hierfür eine Aufgabe mittlerer oder geringer Schwierigkeit gewählt (unter Orientierung an dem Mittelwert des zu messenden Merkmals in der Population), um den Einstieg für die Testperson zu erleichtern. Je mehr Informationen vorab über die Gruppe der Testpersonen vorliegen, umso eher ist bereits für die Einstiegsaufgabe eine treffende Auswahl möglich. Um eine nachfolgende Verzerrung aufgrund dieser A-Priori-Informationen über die Testpersonen zu vermeiden, sollte der adaptive Testalgorithmus allerdings so programmiert sein, dass diese Gruppenzugehörigkeitsinformationen im weiteren Verlauf des Tests nicht mehr zur Aufgabenauswahl herangezogen werden (Thissen & Mislevy, 2000). Für den Schätzprozess der Personenparameter insgesamt spielt die Auswahl der Anfangsaufgabe nur eine untergeordnete Rolle (Frey, 2007).

d) Aufgabenauswahl (How to continue?):
In der meistverwendeten Art des CAT (so genanntes maßgeschneidertes Testen; *tailored testing*) wird nach jeder gegebenen Antwort durch den hinterlegten adaptiven Testalgorithmus ein vorläufiger Personenparameter geschätzt. Diese vorläufige Parameterschätzung bestimmt, welche Aufgabe als nächstes ausgewählt und der Person vorgegeben wird.

Die Schätzung des Personenparameters erfolgt mathematisch in der Regel entweder anhand der sogenannten Maximum Likelihood (ML; z. B. Weighted Likelihood Estimators, WLEs; Warm, 1989; vgl. auch Abschnitt 2.1) oder anhand von bayesianischen Schätzern (z. B. Expected-A-Posteriori-Schätzer, EAP; Bock & Mislevy, 1982). Beide Schätzverfahren eignen sich für die individuelle Personenparameterschätzung, sind jedoch jeweils mit Nachteilen verbunden. Welches Schätzverfahren im Einzelfall vorzuziehen ist, hängt davon ab, ob man mehr Wert auf kleine bedingte Standardfehler (Bayes-Schätzer) oder auf unverzerrte Parameterschätzungen auch in den Extrembereichen der Personenparameterverteilung legt (ML-Schätzer; vgl. Frey, 2007; Segall, 2005). Besteht der CAT pro Person aus etwa 20 Aufgaben oder mehr, führen beide Schätzverfahren zu nahezu identischen Ergebnissen (Thissen & Mislevy, 2000).

Das Vorgehen bei der Aufgabenauswahl lässt sich vereinfacht so beschreiben, dass bei Falschantwort eine einfachere Aufgabe und bei korrekter Antwort eine schwierigere Aufgabe vorgelegt wird. Grundsätzlich haben sich bei der Aufgabenauswahl zwei verschiedene Strategien durchgesetzt (z. B. van der Linden & Pashley, 2010): Entweder wird die folgende Aufgabe so ausgewählt, dass die maximale Menge an Information zum aktuellen (vorläufigen) Personenparameterschätzer gewonnen wird (vgl. Abschnitt 2.1; Hambleton & Cook, 1977; Lord, 1980). Dies bedeutet im Rasch-Modell, dass jeweils die Aufgabe ausgewählt wird, deren Schwierigkeitsparameter *b* dem vorläufigen Personenparameter $\hat{\theta}$ am nächsten kommt und bei der die Lösungswahrscheinlichkeit somit etwa 50 Prozent beträgt (vgl. Frey, 2007). Oder die Aufgabenauswahl wird so getroffen, dass der a posteriori zu erwartende Standardfehler des vorläufigen Personenparameterschätzers und damit dessen Varianz minimiert wird (so genannter Bayes-Ansatz; vgl. Owen, 1975). Da die zugrunde liegenden mathematischen Funktionen dieser beiden Ansätze verwandt sind, führen beide Verfahren in der Regel zur Auswahl derselben Testaufgaben (Weiss & Kingsbury, 1984).

Nachdem eine Aufgabe beantwortet wurde, wird die Schätzung des Personenparameters auf Basis aller vorliegenden Antworten aktualisiert, und die nächste Aufgabe wird ausgewählt. Der Prozess der iterativen Personenparameterschätzung setzt sich so lange fort, bis ein vorab definiertes Abbruchkriterium erreicht wird.

e) Abbruchkriterium (How to stop?):
Üblicherweise wird die Abbruchregel so spezifiziert, dass der Test entweder aufhört, wenn die Personenparameterschätzung hinlänglich präzise ist (d. h. wenn der Standardfehler der Schätzung einen kritischen Wert unterschreitet), wenn eine kritische Anzahl an Testaufgaben vorgegeben wurde oder wenn eine bestimmte Zeitdauer abgelaufen ist. Die erste Variante hat den psychometrischen Vorteil, dass die endgültigen Personenparameterschätzer für alle

Testpersonen eine fast identische Präzision aufweisen. Dies kann jedoch mit interindividuell sehr unterschiedlichen Testlängen verbunden sein (*variable test length*).

Die zweite Variante hat den organisatorischen Vorteil, dass alle Personen gleich viele Aufgaben bearbeiten (*fixed test length*). Allerdings kann die Präzision der endgültigen Personenparameterschätzer zwischen den Testpersonen stärker variieren (Thissen & Mislevy, 2000).

Die dritte Variante hat den organisatorischen Vorteil, dass alle Personen den Test gleichzeitig beenden. Sie ist jedoch für so genannte *Power-Tests* kaum zu empfehlen, in denen die Zeit zur Bearbeitung der Testaufgaben nicht beschränkt wird und nicht in die Auswertung der Testergebnisse eingeht. Handelt es sich um *Speed-Tests*, die die Leistung von Testpersonen unter Zeitdruck erfassen, bietet sich dieses Abbruchkriterium hingegen an (vgl. Thissen & Mislevy, 2000).

Frey (2007) erwähnt als weiteres, offensichtliches Abbruchkriterium den Fall, dass alle Aufgaben des Itempools vorgegeben wurden. Welches Abbruchkriterium gewählt wird oder ob sich eine Kombination aus mehreren Abbruchkriterien empfiehlt, muss in Abhängigkeit der jeweiligen diagnostischen Zielsetzung, des Anwendungskontexts, des Aufgabenpools und der Durchführungsmöglichkeiten entschieden werden.

Es gibt keine generellen Empfehlungen, welche Start-, Prozess- und Abbruchkriterien für einen CAT am besten sind. Die Entscheidung für oder gegen bestimmte Kriterien oder für eine Kombination aus mehreren Kriterien muss im Einzelfall gut durchdacht getroffen werden.

## 2.3 Vor- und Nachteile computerisierten adaptiven Testens

Computerisiertes adaptives Testen ist mit verschiedenen Vor- und Nachteilen verbunden. Der Hauptvorteil dieser Art zu testen liegt in der sehr hohen Messeffizienz, die mit sequentiellen Testalgorithmen nicht erreicht werden kann. *Messeffizienz* ist definiert als Verhältnis von Messpräzision zu Testlänge (Segall, 2005). CAT ermöglicht deutlich kürzere Tests bei derselben Messgenauigkeit oder aber deutlich präzisere Tests bei derselben Testlänge als FIT (für detaillierte Ausführungen zur Messeffizienz siehe Frey, 2007). Zur Verkürzung der Testlänge stellt Weiss (1982) fest, dass im CAT durchschnittlich knapp 50 Prozent weniger Aufgaben vorgegeben werden müssen als im FIT, um dieselbe Messpräzision zu erreichen. Zu ähnlichen Ergebnissen kommen Frey und Ehmke (2007) in einer Echtdaten-Simulationsstudie. De Ayala (2009) spricht sogar von einer Reduktion der Testlänge um 80 Prozent. Hält man die Testlänge konstant, wird die Steigerung der Messpräzision durch CAT bereits aus dem Grundprinzip des adaptiven Testens heraus deutlich: Da die Aufgaben im CAT nach maximaler Information oder der Minimierung des Standardfehlers ausgewählt werden, ergibt sich bei konstanter Testlänge eine präzisere Messung und damit eine höhere Messeffizienz als im FIT. Insbesondere für Personen, die eine extrem hohe oder eine extrem niedrige Merkmalsausprägung haben, entstehen im FIT recht unpräzise Schätzungen der Personenparameter (de Ayala, 2009). Dies liegt daran, dass im FIT nur wenige der vorgegebenen Aufgaben pro Person das Kriterium maximaler Information oder der Minimierung des Standardfehlers erfüllen. Denn in der Regel werden bei diesem Testalgorithmus viele Aufgaben mit mittlerer Schwierigkeit und nur wenige Aufgaben mit extremer Schwierigkeit vorgegeben.

Beim Einsatz multidimensionaler adaptiver Testverfahren (MAT), die zugleich mehrere latente Dimensionen adaptiv messen, ergibt sich eine zusätzliche Steigerung der Messeffizienz, insbesondere wenn die untersuchten Dimensionen hoch miteinander korrelieren. In einer Simulationsstudie analysieren Frey und Seitz (2010) die Messeffizienz unter Bedingungen, die für groß angelegte Vergleichsstudien typisch sind (z. B. wie bei PISA: mehrere, hoch miteinander korrelierte Dimensionen, die simultan gemessen werden). Sie demonstrieren, dass sich die Messeffizienz im MAT gegenüber CAT um das 1.3-fache und gegenüber FIT um das 3.7-fache steigern lässt. In einer Echtdatenstudie, die die Daten von PISA 2000, PISA 2003 und PISA 2006 aus Deutschland verwendet (OECD, 2001, 2004, 2007), finden Frey und Seitz (in press) unter optimalen Bedingungen eine 1.7-fache Steigerung der Messeffizienz im MAT gegenüber FIT.

Vorteilhaft an CAT ist darüber hinaus, dass CAT unter bestimmten Voraussetzungen eine für alle Testpersonen einheitlich präzise Personenparameterschätzung ermöglichen kann, bei der die Standardfehler für alle Personen gleich klein sind. Dies führt zu einer Steigerung der konvergenten Validität von CAT gegenüber FIT, also zu einer Erhöhung der Korrelation von Daten aus Tests, die dasselbe latente Merkmal erfassen (Weiss & Kingsbury, 1984). Eine notwendige Voraussetzung hierfür ist ein Aufgabenpool, der hinreichend groß ist, dass er für alle Merkmalsausprägungen genügend Aufgaben beinhaltet (optimal ist eine Gleichverteilung der Aufgaben über alle in der Personenstichprobe vorhandenen Merkmalsausprägungen; Segall, 2005). Eine weitere Voraussetzung ist die Verwendung eines Abbruchkriteriums, das sich an der Genauigkeit der Parameterschätzung orientiert (Frey, 2007). Von Steigerungen der diskriminanten Validität durch CAT, welche sich durch geringe Korrelationen zwischen Daten aus Tests auszeichnet, die unterschiedliche latente Merkmale erfassen, berichtet Frey (2006, 2007) unter Bezugnahme auf zahlreiche eigene und fremde experimentelle Studien.

Ein weiterer Vorteil von CAT ist, dass die Aufgaben aus dem Aufgabenpool in der Regel länger und besser geheim gehalten werden können als dies im FIT der Fall ist. Die Testpersonen wissen vorab nicht, welche Aufgaben sie tatsächlich bearbeiten müssen, und sie bearbeiten größtenteils unterschiedliche Aufgaben. Um einen bedeutsamen Einfluss der Gedächtnisleistung der Testpersonen auf die eigene oder eine fremde Testleistung zu bewirken, müssten die Testpersonen entweder den gesamten Aufgabenpool auswendig lernen oder der Zufall müsste bemüht werden, so dass die informierte Testperson tatsächlich in großem Maße dieselben Aufgaben zur Bearbeitung erhielte wie die informierende Person (vgl. Wainer, 2000). Beide Varianten können als unwahrscheinlich gelten. Ein Einsatz von CAT in groß angelegten Vergleichsstudien brächte den weiteren Vorteil mit sich, dass das bei solchen Studien in der Regel eingesetzte, hoch komplexe Multi-Matrix-Design und der Druck papierner Testhefte entfielen (Frey & Ehmke, 2007), wodurch Personal- und Materialkosten eingespart werden könnten.

Neben diesen CAT-spezifischen Vorteilen gibt es eine Reihe weiterer Vorteile, die sich aus der Verwendung computerisierter Testverfahren generell ableiten lassen (vgl. Frey, 2007; Linacre, 2000; Parshall, Harmes, Davey & Pashley, 2010; Segall, 2005; Wainer, 2000). Beispielhaft werden nachfolgend einige aufgeführt:

- Es lassen sich Multimedia-Sequenzen in die Aufgabenpräsentation einarbeiten.
- Die Aufgaben können interaktive Elemente enthalten.

- Die Antwortzeit kann millisekundengenau erfasst und zur Auswertung ergänzend hinzugezogen werden.
- Die Antworten können direkt erfasst, gespeichert und auch ohne psychometrisches Fachwissen fehlerfrei bewertet werden.
- Die Aufgabendarstellung kann für Stichproben mit speziellen Bedürfnissen mit wenig Aufwand an individuelle Bedürfnisse angepasst werden (z. B. durch größere Schrift bei Testpersonen mit eingeschränkter Sehfähigkeit).

Computerisiertes adaptives Testen sowie computerisiertes Testen im Allgemeinen ist jedoch auch mit einigen Nachteilen verbunden: Die Güte eines CAT hängt wesentlich von dem Aufgabenpool ab, aus dem die individuell vorzugebenden Aufgaben ausgewählt werden können. Dieser Aufgabenpool sollte im optimalen Falle genügend Aufgaben unterschiedlicher Schwierigkeiten bereitstellen, so dass alle in der Personenstichprobe vorhandenen Ausprägungen des zu messenden Merkmals hinreichend abgedeckt werden. So vorteilhaft CAT in der Anwendung ist, so aufwendig ist die Entwicklung eines solchen Tests (Zara, 1999). Die Aufgaben müssen erstellt beziehungsweise gegebenenfalls aus einem bestehenden FIT adaptiert und kalibriert werden. Dies ist ein sehr zeitaufwendiger Prozess, der umfangreiches Expertenwissen in der zu messenden inhaltlichen Domäne und psychometrisches Wissen, insbesondere eine Versiertheit im Bereich der Item-Response-Theorie, erfordert. Dies dürfte ein Hauptgrund dafür sein, weshalb es im deutschsprachigen Raum noch immer nur relativ wenige adaptive Testverfahren gibt (vgl. Frey, 2007). Welche einzelnen inhaltlichen, psychometrischen, organisatorischen und politischen Schritte und Überlegungen bei der Entwicklung eines CAT zu bedenken sind, stellt Zara (1999) überblicksartig dar und weist zugleich darauf hin, dass sich der hohe Aufwand im Nachhinein jedoch bezahlt machen kann. Auch Olson (2003) schreibt, dass die Kosten für die Entwicklung eines hochwertigen CAT in etwa vergleichbar seien mit denen für die Entwicklung eines hochwertigen Papier-Bleistift-Tests, dass der Einsatz und die Pflege eines fertigen CAT jedoch deutlich ökonomischer ausfielen als für einen herkömmlichen Test.

Nachteilig ist außerdem, dass die Aufgaben im CAT je nach verwendeter Software an ein Multiple-Choice-Antwortformat gebunden sind, da sie ad hoc auswertbar sein müssen. Zenisky und Sireci (2002) weisen jedoch auf die Forschung zu automatisiertem Scoring (d. h. zur computergestützten, verzögerungsarmen Bewertung von Antworten) und auf die Weiterentwicklung computerisierter und adaptiver Testprogramme hin. Eine Erweiterung des verwendbaren Antwortformats von reinen Multiple-Choice-Formaten hin zu halboffenen oder offenen Formaten im CAT könnte in Zukunft umsetzbar sein (vgl. auch Olson, 2003). Eine Einbindung von Aufgaben mit offenem Antwortformat im CAT kann nützlich sein, auch wenn die Antworten auf diese Aufgaben nicht in die Parameterschätzung eingehen: Da die Aufgaben im CAT anhand des vorläufigen Personenparameterschätzers ausgewählt werden, beinhalten sie im Mittel mehr Informationen über die Merkmalsausprägung einer Person als Aufgaben mit offenem Antwortformat im FIT (vgl. Frey & Seitz, in press).

Ein anderer Nachteil, der sich auf computerisiertes Testen im Allgemeinen bezieht, ist, dass die notwendige Hardware zur Verfügung stehen muss. Im Bereich schulischer Leistungstests zeigt sich beispielsweise, dass sich zwar die Computerausstattung in den Schulen Deutschlands in den vergangenen Jahren deutlich verbessert hat (Krützer & Probst, 2006), eine vollständige Umstellung

von Schulleistungsstudien auf eine computerisierte Durchführung jedoch noch nicht möglich ist. Die technische Ausstattung zwischen den Schulen in Deutschland und erst recht zwischen verschiedenen Teilnehmerstaaten internationaler Schulleistungsstudien unterscheidet sich noch stark und würde die Objektivität und die Vergleichbarkeit der Ergebnisse beeinträchtigen. Die ungleichen Hardware-Voraussetzungen sind vermutlich momentan noch ein Hauptnachteil, der mit computerisiertem Testen im Allgemeinen und CAT im Besonderen verbunden ist. Wie groß der Effekt ungleicher technischer Voraussetzungen auf die Testleistung tatsächlich ist, ist wissenschaftlich noch nicht hinreichend geklärt und wird laut Segall (2005) in den kommenden Jahren verstärkt Gegenstand der Forschung sein. Die unterschiedliche individuelle Vertrautheit mit Computern und weitere personenbezogene Störvariablen, die nicht mit dem im Test zu erfassenden Merkmal in Verbindung stehen, mögen die Fairness und die Validität der Ergebnisse computerisierter Tests zusätzlich beeinträchtigen (Frey, 2007; Linacre, 2000).

Empirische Befunde zur Motivation zur Testbearbeitung im CAT (vgl. Abschnitt 4.2) lassen bislang keine eindeutigen Schlüsse zu, ob die Testfairness und/oder die Validität der Ergebnisse im Sinne einer Messung der maximalen Leistung im CAT beeinträchtigt sind. Die Behauptung von Wainer (2000) in Bezug auf die Motivationslage der Testpersonen in einem CAT („Each individual stays busy productively – everyone is challenged but not discouraged"; S. 11) ist vor dem Hintergrund der aktuellen empirischen Befundlage nicht haltbar. Es ist ein Ziel der vorliegenden Arbeit, diese Behauptung empirisch-experimentell und mit der nötigen Sorgfalt zu prüfen.

## 2.4 Anwendungen computerisierten adaptiven Testens in der empirischen Bildungsforschung

Angesichts der genannten zahlreichen Vorteile von CAT stellt diese Art zu testen nicht nur im allgemeinen, sondern speziell auch im Bereich der empirischen Bildungsforschung eine vielversprechende Alternative zu FIT dar. Über die Möglichkeiten und Grenzen eines Einsatzes von CAT in groß angelegten Vergleichsstudien wird in der wissenschaftlichen Gemeinschaft gegenwärtig viel diskutiert (vgl. z. B. http://www.ctb.com/iacat). Nichtsdestotrotz findet CAT derzeit, insbesondere im deutschsprachigen Raum, erst eingeschränkt Anwendung. Ein Grund mag sein, dass eine gute Kenntnis der IRT unerlässlich ist, um einen CAT zu entwickeln und zu verstehen. Die benötigten psychometrischen Kenntnisse gehen möglicherweise über das Maß hinaus, in welchem IRT vielerorts an den Universitäten in Deutschland gelehrt wird. Es gibt jedoch einige internationale und nationale Beispiele für eine erfolgreiche Nutzung dieses Testalgorithmus, die im Folgenden kurz aufgeführt werden.

Im US-amerikanischen Raum zählt die *Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery* (CAT-ASVAB; Moreno, 1997) des Verteidigungsministeriums zu den erfolgreich implementierten und bekanntesten adaptiven Testverfahren. Die CAT-ASVAB wird im US-Militär zur Personalauswahl eingesetzt. Im US-Bildungsbereich wird kontrovers über eine Nutzung von CAT zur ökonomischen Durchführung der seit dem *No Child Left Behind Act* (NCLB) 2001 verbindlichen flächendeckenden Schulleistungstests diskutiert (Harmon, 2010; Trotter, 2003). Die Kontroverse macht deutlich, dass ein fehlendes oder nicht ausreichendes Verständnis der Funktionsweise von CAT zu Misstrauen gegenüber dieser Art zu testen führen kann (Trotter, 2003).

Seitens des US-Bildungsministeriums bestehen Bedenken bezüglich eines „out-of-level"-Testens: Die Schulleistungstests dienen der Überprüfung inhaltlicher, klassenstufenspezifischer Standards, weswegen die US-Regierung eine Beschränkung der Testaufgaben auf diese stufenspezifischen Inhalte verlangt (Harmon, 2010). Aus psychometrischer Sicht mindert dies zwar die Qualität des CAT, da dieser gerade für leistungsstarke und leistungsschwache Schulkinder präzisere Personenparameterschätzer bereitstellen könnte, indem Aufgaben einer höheren oder niedrigeren Klassenstufe verwendet werden (Way, 2010). Dennoch lässt sich dieses Dilemma leicht lösen, indem man den Aufgabenpool auf klassenstufenspezifische Aufgaben beschränkt. In einem Bundesstaat (Oregon) kommt CAT bereits im Rahmen von NCLB zum Einsatz (Harmon, 2010; Kingsbury & Hauser, 2004).

Ein weiteres Beispiel aus dem Bereich der empirischen Bildungsforschung ist die Entwicklung und Implementation eines längsschnittlich angelegten CAT in Portland, Oregon (USA), mit dem die Leistung von Schulkindern über mehrere Jahre hinweg erfasst und kontrolliert wird (Kingsbury & Houser, 1999). Ferner setzen mehrere US-amerikanische Bundesstaaten CAT ein, um die Kompetenz ihrer Schülerinnen und Schüler regelmäßig zu messen (Olson, 2003). Auch die zwei größten Eignungstests in den USA zur Auswahl von Studierenden betriebswirtschaftlicher Fächer (*Graduate Management Admission Test*, GMAT; durchgeführt von Pearson Vue) und von Doktorandinnen und Doktoranden verschiedenster Fächer (*Graduate Record Examniation*, GRE; durchgeführt vom *Educational Testing Service*, ETS; Mills & Steffen, 2000) werden als CAT angeboten. Wie die Organisation für Wirtschaftliche Zusammenarbeit und Entwicklung in einer Broschüre über die internationale Schulleistungsstudie PISA schreibt (OECD, 2006), ist vorgesehen, ab 2012 auch hier CAT-Teile einzubauen und zu erproben. Dies könnte gerade angesichts der großen Heterogenität von Schülerleistungen in den Teilnehmerstaaten präzisere Kompetenzschätzungen ermöglichen. Allerdings deuten die Ergebnisse von Frey, Seitz und Kröhne (2010) darauf hin, dass der Messeffizienz-Gewinn und in Folge dessen der Nutzen einer Umstellung von FIT auf MAT bei bestehenden Studien wie PISA geringer ausfallen könnte als bei neu entwickelten Kompetenztests, die von Beginn an adaptiv konzipiert werden.

In Deutschland wird der Einsatz adaptiven Testens im Bereich der empirischen Bildungsforschung diskutiert, doch bisher kaum durchgeführt. Eines der wenigen Beispiele einer regelmäßigen Nutzung von CAT an größeren Stichproben in Deutschland stellt das so genannte Computer-Assistierte Testen im Rahmen eignungsdiagnostischer Untersuchungen der Bundeswehr dar, das Tests zu abstrakt-logischem Denken, verbal-logischem Denken und Rechenfähigkeit beinhaltet (Pawlik, 1997; Steyer & Partchev, 2000). Frey und Ehmke (2007) analysieren in einer Simulationsstudie Möglichkeiten und Hindernisse eines flächendeckenden Einsatzes von CAT zur Überprüfung der Bildungsstandards. Sie bescheinigen viele Vorteile, diese regelmäßigen Erhebungen computerisiert und adaptiv durchzuführen, weisen jedoch auf einige Punkte hin, die vorab geklärt werden müssten. Unter anderem steht die Klärung der Frage aus, ob CAT zu einer differentiellen, von Personenmerkmalen abhängigen Beeinflussung der Motivation zur Testbearbeitung führt, wodurch die Validität der Testergebnisinterpretation und/oder die Testfairness geschmälert würde/n. Die vorliegende Arbeit trägt zur Klärung dieser Frage bei, indem die Auswirkungen von CAT auf die Motivation zur Testbearbeitung unter Berücksichtigung weiterer Test- und Personenmerkmale differenziert experimentell untersucht werden.

# 3 Theorien zur Leistungsmotivation

Nachdem in Abschnitt 2 die Grundzüge computerisierten adaptiven Testens, einige Vor- und Nachteile dieser Art zu testen sowie einige Anwendungsbeispiele vorgestellt wurden, folgt in diesem Abschnitt eine Darstellung der motivationstheoretischen Grundlagen der Arbeit. Da die Bearbeitung eines Leistungstests eine typische Leistungssituation repräsentiert, beschränkt sich die Darstellung der motivationstheoretischen Grundlagen auf den Bereich der Leistungsmotivation. Auf Basis der theoretischen Grundlagen zu CAT im vorherigen Abschnitt und zur Leistungsmotivation in diesem Abschnitt werden in Abschnitt 5 die Hypothesen abgeleitet.

Nach einer begrifflichen Annäherung an die Konstrukte Leistungsmotivation und Motivation zur Testbearbeitung sowie einer generellen Betrachtung, wie Verhalten in Leistungssituationen entsteht (Abschnitt 3.1), werden in Abschnitt 3.2 die für die vorliegende Arbeit bedeutsamen Theorien und Modelle vertieft dargestellt. Zunächst wird mit dem Risikowahl-Modell nach Atkinson (1957, 1964) die klassische Leistungsmotivationstheorie erläutert und kritisch diskutiert (Abschnitt 3.2.1). Dieses Modell liegt einigen Studien zur Motivation zur Testbearbeitung im CAT zugrunde und ist für das Verständnis der kritischen Auseinandersetzung mit diesen Studien wichtig (vgl. Abschnitt 4). Zudem bildet das Risikowahl-Modell die Grundlage für das Erwartung-Wert-Modell der Leistungsmotivation nach Eccles und Wigfield (2002; Abschnitt 3.2.2), welches das theoretische Fundament der vorliegenden Arbeit ist.

## 3.1 Begriffsdefinitionen und Grundüberlegungen zu leistungsmotiviertem Verhalten

*Motivation* (von lateinisch *movere*: bewegen) kann allgemein definiert werden als Energetisierung und aktivierende Ausrichtung auf einen positiv bewerteten Zielzustand (z. B. Rheinberg, 2008). Diese Definition deutet an, dass Motivation neben einem statischen Aspekt der Bereitschaft einen dynamischen Aspekt enthält (Schiefele, 1996). Dieser kann als Prozess beschrieben werden, durch den das zielgerichtete Verhalten ausgelöst und aufrechterhalten wird (Heckhausen & Heckhausen, 2006; Prochaska, 1998; Rudolph, 2009; Schunk, Pintrich & Meece, 2008). Dieser Prozess und damit die Erklärung der Genese aktueller Motivation ist eine Grundlage der Erwartung-Wert-Modelle der Motivation, auf die in Abschnitt 3.2 eingegangen wird (Schiefele, 1996, 2009).

Die allgemeine Motivationsdefinition ist auf unterschiedliche Inhaltsbereiche anwendbar und muss daher für den Leistungskontext konkretisiert werden. Dies erscheint auch angesichts der zahlreichen Definitionen und Annäherungen an das Konstrukt der Leistungsmotivation unerlässlich (vgl. Metaanalyse von Murphy & Alexander, 2000). Von *Leistungsmotivation* spricht man, wenn es um die Auseinandersetzung mit einem individuell gesetzten Güte- oder Tüchtigkeitsmaßstab geht (Brunstein & Heckhausen, 2006). Beschränkt man sich auf die Betrachtung von Motivation in einer Leistungstestsituation, lässt sich *Motivation zur Testbearbeitung* als Spezialfall der Leistungsmotivation spezifizieren. Diese spezifische Motivation ist in Anlehnung an Baumert und Demmrich (2001) definiert als Bereitschaft, Testaufgaben unter Anstrengung und bestmöglich zu bearbeiten und diese Anstrengung über die Testdauer hinweg aufrechtzuerhalten. Der statische

Aspekt der allgemeinen Motivationsdefinition wird hier durch den Begriff „Bereitschaft“ ausgedrückt: Die Person ist auf die erfolgreiche Bearbeitung des Tests ausgerichtet. Damit die Person als motiviert bezeichnet werden kann, muss die erfolgreiche, bestmögliche Bearbeitung des Tests positiv konnotiert und als Ziel anerkannt sein (vgl. auch Wise & DeMars, 2005). Hinzu kommt eine individuelle Erwartung der Person, anhand derer sie ihr Leistungsverhalten im Test und ihr Testergebnis misst und bewertet (Güte-/Tüchtigkeitsmaßstab). Der dynamische Aspekt der Motivationsdefinition wird in der Persistenz der Anstrengungsbereitschaft deutlich, die durch die Leistungsorientiertheit der Testsituation initiiert wird und sich auf die gesamte Dauer der Testsituation bezieht. Damit Motivation zur Testbearbeitung und schließlich Leistungsverhalten im Test entsteht, muss also ein in der Person liegendes, leistungsorientiertes Merkmal existieren (Güte-/Tüchtigkeitsmaßstab), und es muss ein leistungsorientierter Kontext vorliegen (Testsituation).

Der prozesshafte Charakter der Motivation zur Testbearbeitung lässt sich in einem Flussmodell bildlich darstellen (Abbildung 3.1; in Anlehnung an Frey, 2006; Heckhausen & Heckhausen, 2006; Rheinberg, 2008). Das *Grundmodell motivierten Verhaltens in einer Testsituation* bildet die Notwendigkeit leistungsorientierter Personen- und Situationsmerkmale für das Zustandekommen leistungsmotivierten Verhaltens in einem Test ab. Eine weitere Einflussgröße auf die Motivation zur Testbearbeitung können Testmerkmale darstellen. Die Berücksichtigung von Testmerkmalen in dem Modell kann dazu beitragen, die Motivation und das Leistungsverhalten speziell in Testsituationen genauer vorherzusagen und so dem Missstand entgegenzuwirken, dass bis heute keine „präzise Theorie [existiert], die erklärt, wie Leistungsmotivation auf die einzelnen Schritte der Aufgabenbearbeitung […] einwirkt“ (Brunstein & Heckhausen, 2006, S. 171).

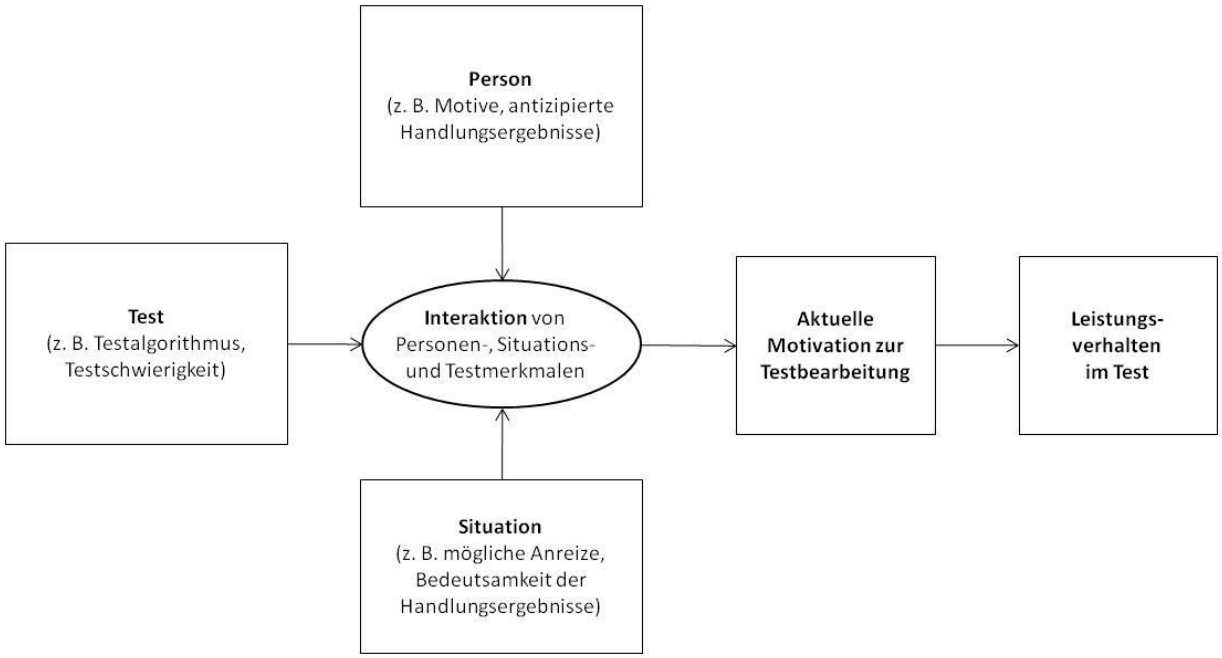



*Abbildung 3.1.* Grundmodell motivierten Verhaltens in einer Testsituation (in Anlehnung an Frey, 2006; Heckhausen & Heckhausen, 2006; Rheinberg, 2008).

Unter *Personenmerkmalen* sind in dem Modell relativ zeitstabile, überdauernde Aspekte zusammengefasst, die situationsübergreifend stets in ähnlicher Weise auf das Leistungsverhalten einer Person einwirken. Dabei baut das Grundmodell nicht auf einer speziellen Theorie auf, sondern ist mit unterschiedlichen Leistungsmotivationstheorien vereinbar. Personenmerkmale können daher inhaltlich unterschiedlich ausgefüllt werden, zum Beispiel in Form von universellen Bedürfnissen (Murray, 1938), impliziten oder expliziten Motiven (Brunstein, 2006; Brunstein & Hoyer, 2002; McClelland, Koestner & Weinberger, 1989) oder eines individuellen Gütemaßstabs (Brunstein & Heckhausen, 2006). Auch die Antizipation von Handlungsergebnissen ist als Personenmerkmal aufzufassen. Dies wurde in frühen Modellen zur Erklärung leistungsmotivierten Verhaltens nicht berücksichtigt. Ältere Ansätze betrachteten Leistungsmotivation aus einer behavioristischen Perspektive und gingen von einer reinen Reiz-Reaktions-Verbindung aus (Stimulus-Response-Modelle). Der Mensch wurde als unreflektiert agierendes Wesen eingestuft (z. B. Watson, 1913). Später entwickelte motivationstheoretische Ansätze beziehen vermittelnde und Handlungsfolgen antizipierende kognitive und emotionale Aspekte ausdrücklich mit ein (vgl. Stimulus-Organism-Response-Consequence-Modelle; z. B. Kanfer, 1987). Diese Erweiterung ist für die vorliegende Arbeit zentral bedeutsam, da die Ausprägung der leistungsbezogenen Personenmerkmale in Interaktion mit Merkmalen des Tests (z. B. mit der Testschwierigkeit) das Testverhalten maßgeblich beeinflussen kann (vgl. Abschnitte 3.2, 4 und 5).

Trotz dieser überdauernden Personenmerkmale verhält sich eine Person nicht in jeder Leistungssituation gleich. Vielmehr hängt das jeweilige Verhalten zugleich ab von den *Situationsmerkmalen*. Leistungsmotiviertes Verhalten bedarf, im Gegensatz zu ausschließlich triebgesteuertem Verhalten, eines situationsspezifischen Auslösers und entsteht nicht „von alleine“ aus einer Person heraus (Rheinberg, 2008). Situationsmerkmale sind situationsspezifische intrinsische und extrinsische Anreize, die Aufforderungscharakter haben (Heckhausen & Heckhausen, 2006). Diese Anreize können an die Handlungstätigkeit selbst, an das Handlungsergebnis oder an die Bedeutsamkeit des Handlungsergebnisses und damit an verschiedene Arten von Handlungsfolgen geknüpft sein und sind für sich genommen zunächst unabhängig von den antizipierten persönlichen Erfolgserwartungen. Ein situationsspezifischer Hinweisreiz zum Auslösen leistungsmotivierten Verhaltens mag ein Klassenraum mit Einzeltischen sein, auf denen Aufgabenzettel zur Mathematik liegen. Ein typischer extrinsischer Anreiz für ein gutes Abschneiden in einem Test könnte sein, mit dem Testergebnis ein gewisses Ziel, beispielsweise die Zulassung zu einem Studiengang, zu erreichen. Ist das Testergebnis, wie in diesem Beispiel, mit individuellen Konsequenzen verbunden, handelt es sich bei der Leistungssituation um eine *High-Stakes*-Testsituation (im Gegensatz zu *Low-Stakes*-Testsituationen, die keinerlei persönliche Konsequenzen für die Testperson haben; vgl. Abschnitt 4).

Unter *Testmerkmalen* schließlich werden verschiedene Spezifika eines Leistungstests verstanden, die für sich genommen sowie im Zusammenspiel mit Personen- und Situationsmerkmalen einen Einfluss auf die aktuelle Motivation zur Testbearbeitung haben können. So mögen Motivation und Verhalten von Personen in einem Test, der Aufgaben verschiedenster Schwierigkeiten beinhaltet, anders ausgeprägt sein, als in einem Test, der lediglich aus sehr schwierigen Aufgaben besteht. Diese Motivations- und Verhaltensunterschiede zwischen den Personen in den beiden Schwierigkeitsbedingungen können darüber hinaus beispielsweise in Abhängigkeit von den individuellen Gütemaßstäben der Personen differieren (z. B. Rink, 1994). Testmerkmale und ihre Interaktionen mit Personen- und Situationsmerkmalen stehen im Mittelpunkt der vorliegenden Arbeit, da sie zu differentiellen Effekten auf die aktuelle Motivation zur

Testbearbeitung führen können. Sind die Effekte differentiell, können sie die Testfairness und/oder die Validität der Testergebnisinterpretation beeinträchtigen (vgl. nähere theoretische Ausführungen in Abschnitt 3.2.2, empirische Befunde in Abschnitt 4 und Hypothesen in Abschnitt 5).

In dem Grundmodell motivierten Verhaltens in einer Testsituation werden verschiedene Annahmen deutlich: Leistungsmotiviertes Verhalten in einem Leistungstest wird durch die aktuelle Motivation zur Testbearbeitung unmittelbar beeinflusst, während die Personen-, Situations- und Testmerkmale als indirekte Einflussgrößen des Verhaltens aufzufassen sind. Das leistungsmotivierte Verhalten lässt sich sowohl auf eine überdauernde stabile Personeneigenschaft (in der Motivationsliteratur in der Regel als „Leistungsmotiv“ oder *Trait*-Motivation bezeichnet) als auch auf eine situationsspezifische instabile Komponente (in der Motivationsliteratur als „Leistungsmotivation“ oder *State*-Motivation bezeichnet; vgl. Prochaska, 1998) zurückführen. Beiden Aspekten der Leistungsmotivation sollte bei der Erklärung leistungsmotivierten Verhaltens Beachtung geschenkt werden.

Abgeleitet aus diesen begrifflichen und modelltheoretischen Vorüberlegungen lässt sich zusammenfassen, dass (a) es bislang keine fundierte Theorie gibt, die die Wirkung von Leistungsmotivation in Bezug auf die Testbearbeitung erklärt, (b) bei der Analyse der Motivation zur Testbearbeitung Personen-, Situations- und Testmerkmale zu berücksichtigen sind, und (c) im Rahmen einer solchen Analyse sowohl stabile (Trait-Motivation) als auch situationsspezifische (State-Motivation) Motivationskomponenten einzubeziehen sind. Diese Erkenntnisse werden im Verlauf der weiteren Arbeit wiederholt aufgegriffen und sind für die Ableitung der Fragestellungen und Hypothesen von zentraler Bedeutung (vgl. Abschnitt 5). Um ein differenzierteres Bild der Entstehung aktueller Leistungsmotivation zu erhalten, werden im folgenden Abschnitt Erwartung-Wert-Modelle vorgestellt, deren Typ nahezu alle aktuellen Leistungsmotivationstheorien in ihren Grundzügen entsprechen (Beckmann & Heckhausen, 2006; Eccles & Wigfield, 1995).

## 3.2 Erwartung-Wert-Modelle der Leistungsmotivation

Erwartung-Wert-Modelle haben sich zur Vorhersage von Leistungsverhalten empirisch gut bewährt. Ein wesentliches, positiv hervorzuhebendes Charakteristikum der meisten Erwartung-Wert-Modelle ist, dass sie sowohl Personen- als auch Situationsmerkmale berücksichtigen und Trait- und State-Aspekte der Leistungsmotivation zugleich analysiert werden können (Schmalt & Langens, 2009; vgl. Ausführungen zum Grundmodell motivierten Verhaltens in einer Testsituation in Abschnitt 3.1). Das Risikowahl-Modell von Atkinson (1957, 1964) stellt eines der ersten Erwartung-Wert-Modelle der Leistungsmotivation dar. Zugleich gilt es heute als die „klassische“ Leistungsmotivationstheorie.

Im Folgenden wird das Risikowahl-Modell von Atkinson (1957, 1964) erläutert (Abschnitt 3.2.1). Nach einer kritischen Auseinandersetzung mit dem Modell aus heutiger Sicht wird mit dem Erwartung-Wert-Modell der Leistungsmotivation von Eccles und Wigfield (2002) eine Weiterentwicklung des Risikowahl-Modells vorgestellt (Abschnitt 3.2.2). Dieses Modell wird zunächst in seiner Komplexität beschrieben. Anschließend erfolgt eine Fokussierung auf den für die

Vorhersage von Motivation zur Testbearbeitung besonders relevanten Teil des Modells, auf Basis dessen in Abschnitt 5 die Hypothesen dieser Arbeit abgeleitet werden.

### 3.2.1 Das Risikowahl-Modell von Atkinson (1957, 1964)

Das Risikowahl-Modell gilt als die klassische Leistungsmotivationstheorie, die trotz einiger Schwächen bis heute viele Erwartung-Wert-Modelle der Leistungsmotivation prägt und vielfältige Forschung anregt (vgl. Überblick in Kuhl, 1983; Schunk et al., 2008). Es gelang erstmals Atkinson (1957, 1964), dispositionelle und situative Komponenten der Leistungsmotivation gleichwertig in ein Modell zu integrieren (Beckmann & Heckhausen, 2006; Funder, 2006; Schmalt & Langens, 2009). Im Folgenden werden diese Komponenten sowie das Gesamtmodell kurz geschildert, da das Risikowahl-Modell die Grundlage einiger der wenigen bestehenden empirischen Studien zur Motivation zur Testbearbeitung in FIT und CAT bildet. Eine Kenntnis des Modells ist für das Verständnis der kritischen Auseinandersetzung mit diesen Arbeiten erforderlich (Abschnitt 4.3). Zudem hilft sie, die Besonderheiten des Erwartung-Wert-Modells der Leistungsmotivation von Eccles und Wigfield (2002) zu verstehen, das der vorliegenden Arbeit zugrunde liegt.

#### *3.2.1.1 Personenkomponente: Leistungsmotiv*

Motive sind latente Merkmale, die nicht direkt beobachtbar sind und die in spezifischen, das Motiv anregenden Situationen (d. h. beim Leistungsmotiv: in Leistungssituationen) verhaltenswirksam werden. Motive gelten als zeitlich stabile Persönlichkeitseigenschaften, wobei Weiner (1994) darauf hinweist, dass das Ausmaß der situationsübergreifenden Stabilität des Leistungsmotivs nicht zufriedenstellend geklärt ist.

Atkinson (1957, 1964) bezeichnet das Leistungsmotiv als überdauernde Persönlichkeitseigenschaft, die sich in die beiden voneinander unabhängigen Komponenten *Hoffnung auf Erfolg* (Erfolgsmotiv; $M_e$) und *Furcht vor Misserfolg* (Misserfolgsmotiv; $M_m$) gliedert (vgl. auch McClelland, Atkinson, Clark & Lowell, 1953; Schunk et al., 2008). Er schreibt diesen beiden Motivkomponenten eine starke affektive Prägung zu, indem er sie über emotionale Antizipationen definiert. So definiert er das Erfolgsmotiv als Eigenschaft, Erfolg aufzusuchen und damit positiven leistungsbezogenen Affekt wie Stolz oder Freude zu empfinden und zu maximieren ("a capacity for taking pride in accomplishment“; Atkinson, 1964, S. 241). Das Misserfolgsmotiv hingegen versteht er entsprechend als Disposition, Misserfolg zu vermeiden, um so negativen leistungsbezogenen Affekt wie Scham oder Ärger zu minimieren oder ihm ganz auszuweichen. Die Motivausprägungen gelten als erlernt: Hat eine Person in der Vergangenheit die Erfahrung gemacht, herausfordernde Aufgaben meistern zu können und dabei Stolz zu empfinden, wird sie dieses Gefühl auch in zukünftigen ähnlichen leistungsthematischen Situationen antizipieren und sich den Aufgaben stellen – dies allerdings nur, wenn der antizipierte emotionale Zustand nach Lösung der Aufgabe positiver ist als der aktuelle emotionale Zustand. Andernfalls bestünde keine Notwendigkeit, tätig zu werden und aus dem jetzigen Zustand herauszutreten. Umgekehrt wird eine Person, die sich an vorwiegend negative leistungsbezogene Erfahrungen erinnert, entsprechende Situationen wenn möglich voraussichtlich meiden.

Die Unterteilung des dispositionellen Leistungsmotivs in die beiden voneinander unabhängigen Komponenten Hoffnung auf Erfolg und Furcht vor Misserfolg ist eine der zentralen Grundannahmen

des Risikowahl-Modells (Schmalt & Langens, 2009). Die separate Betrachtung der beiden Komponenten hat sich bei der Messung von Leistungsmotiven durchgesetzt (Lang & Fries, 2006; Schmalt, 1996). Hinsichtlich der Zugänglichkeit dieser Motivkomponenten für das Bewusstsein und in Folge dessen hinsichtlich der Frage, ob die Motivkomponenten implizit, zum Beispiel projektiv, oder respondent erfasst werden sollen, besteht jedoch Uneinigkeit (Brunstein, 2006; Brunstein & Hoyer, 2002; vgl. auch McClelland et al., 1989; Spangler, 1992).

#### *3.2.1.2 Situationskomponenten: Erwartung und Wert*

Bezüglich der situativen Komponenten des Risikowahl-Modells unterscheidet Atkinson (1964) vier Parameter: Die *Erfolgserwartung* $W_e$ drückt aus, wie hoch eine Person die Wahrscheinlichkeit einschätzt, eine bestimmte vor ihr liegende, konkrete Aufgabe erfolgreich zu bearbeiten. Diese subjektive Erfolgswahrscheinlichkeit ergibt sich aus den Erfahrungen, die die Person in früheren, ähnlichen Leistungssituationen oder mit ähnlichem Aufgabenmaterial gemacht hat, und ist von der wahrgenommenen Schwierigkeit S der zu bearbeitenden Aufgaben abhängig (Weiner, 1994): $W_e = 1-S$. Je schwieriger eine Aufgabe wahrgenommen wird, umso geringer ist die Erfolgserwartung; je einfacher eine Aufgabe wahrgenommen wird, umso höher ist die Erfolgserwartung. Geht man davon aus, dass eine Leistungssituation, zum Beispiel in Form einer Aufgabenbearbeitung, lediglich in die beiden alternativen Ereignisse Erfolg oder Misserfolg münden kann, ergibt sich die *Misserfolgserwartung* $W_m$ als Komplementärwahrscheinlichkeit von $W_e$: $W_m = 1-W_e$ (Atkinson, 1964).

Die anderen beiden situativen Komponenten des Modells sind der Erfolgs- und der Misserfolgsanreiz. Der *Erfolgsanreiz* $A_e$ bezieht sich darauf, wie attraktiv eine Person einen Erfolg bei einer konkreten Aufgabe einschätzt, das heißt, welchen Wert es für sie hat, bei dieser Aufgabe zu reüssieren. Atkinson (1964) schreibt dieser Komponente sowie der komplementären Komponente *Misserfolgsanreiz* $A_m$ inhaltlich eine kognitiv-affektive Prägung zu, indem er den Anreiz abhängig macht von dem antizipierten Stolz bei Erfolg (Erfolgsanreiz $A_e$) beziehungsweise von der antizipierten Scham bei Misserfolg (Misserfolgsanreiz $A_m$). Atkinson setzt den Erfolgsanreiz explizit mit der wahrgenommenen Aufgabenschwierigkeit gleich. Je schwieriger eine Person eine Aufgabe einschätzt, umso mehr Stolz wird bei Lösung der Aufgabe empfunden und umso attraktiver ist es für die Person, die Aufgabe zu lösen. Die wahrgenommene Aufgabenschwierigkeit definiert Atkinson dabei als komplementär zu der Erfolgserwartung, so dass sich als weitere zentrale Annahme des Risikowahl-Modells folgende formale Verknüpfung ergibt: $A_e = S = 1-W_e$. Damit ergibt sich zwingend auch, dass Atkinson die Aufgabenschwierigkeit und den Erfolgsanreiz gleichsetzt mit der Misserfolgserwartung. Im Bezug auf den Misserfolgsanreiz nimmt Atkinson hingegen eine etwas andere formale Definition vor: Da sich der Misserfolgsanreiz auf ein „schockartiges“, aversives Ereignis bezieht (Atkinson, 1964), definiert Atkinson ihn als negative Erfolgswahrscheinlichkeit: $A_m = -W_e$. Je einfacher also eine Aufgabe erscheint und je höher dementsprechend die Erfolgserwartung ist, umso beschämender ist ein Misserfolg bei dieser Aufgabe und umso unattraktiver ist es, die Aufgabe zu lösen.

Aus den vorgestellten Formeln und inhaltlichen Definitionen wird deutlich, dass die situationsspezifischen Parameter, Erwartung und Anreiz, im Risikowahl-Modell per definitionem vollständig voneinander abhängig und negativ-linear miteinander verknüpft sind. Erfolgs- und Misserfolgsanreiz sind eindeutig und ausschließlich durch die subjektive Erfolgswahrscheinlichkeit determiniert (Schmalt & Langens, 2009).

### *3.2.1.3 Das Gesamtmodell*

Das Gesamtmodell spezifiziert die Interaktionen zwischen den in den vorherigen Abschnitten vorgestellten Personen- und Situationskomponenten. Atkinson (1964) nimmt zwei so genannte motivationale Tendenzen an: die Tendenz, Erfolg aufzusuchen (*Erfolgstendenz* $T_e$; auch als Annäherungstendenz bezeichnet), und die Tendenz, Misserfolg zu vermeiden (*Misserfolgstendenz* $T_m$; auch als Vermeidungstendenz bezeichnet). Zur Berechnung der beiden motivationalen Tendenzen multipliziert er die jeweiligen Personen- und Situationskomponenten miteinander, so dass sich ergibt: $T_e = M_e \times W_e \times A_e$ als Erfolgstendenz und $T_m = M_m \times W_m \times A_m$ als Misserfolgstendenz.

In Bezug auf die Erfolgstendenz ($T_e$) weist Atkinson auf folgende Implikationen seines Modells hin: Die Erfolgstendenz ist bei Aufgaben mittlerer wahrgenommener Schwierigkeit und damit bei mittlerer Erfolgserwartung am höchsten, da dann das Produkt aus Erwartung und Wert maximal ist. Die relative Bedeutung des Erfolgsmotivs für die Erfolgstendenz ist dabei gegenüber den situativen Komponenten umso geringer, je stärker die Erfolgswahrscheinlichkeit von 50 Prozent abweicht, je extremer also die wahrgenommenen Aufgabenschwierigkeiten sind. Bei Aufgaben mittlerer wahrgenommener Schwierigkeit ist der Einfluss des überdauernden Erfolgsmotivs auf die Erfolgstendenz maximal. Dies bedeutet andersherum auch, dass insbesondere Personen mit hohem Erfolgsmotiv eine Präferenz für Aufgaben mittlerer wahrgenommener Schwierigkeit zeigen sollten (Atkinson, 1964).

Für die Misserfolgstendenz ($T_m$) formuliert Atkinson (1964) sehr ähnliche Implikationen wie für die Erfolgstendenz: Ist das Misserfolgsmotiv relativ stark, sollten insbesondere Aufgaben mittlerer wahrgenommener Schwierigkeit, also mittlerer Erfolgserwartung, zu einer hohen Misserfolgstendenz führen. Bei diesen Aufgaben sollte zudem die relative Bedeutung des Misserfolgsmotivs für die Misserfolgstendenz im Vergleich zu situativen Einflussfaktoren maximal sein. Hieraus lässt sich schließen, dass insbesondere Personen mit hohem Misserfolgsmotiv, das heißt, mit hoher Testängstlichkeit (Atkinson, 1964), Aufgaben mittlerer wahrgenommener Schwierigkeit meiden. Stattdessen sollten sie, sofern sie die Leistungssituation nicht verlassen können, Aufgaben präferieren, die sie als sehr einfach oder sehr schwierig wahrnehmen.

Aus diesen Implikationen des Risikowahl-Modells lassen sich interessante Annahmen über die Motivation zur Testbearbeitung im CAT ableiten. Dies gilt zumindest, wenn die objektive Erfolgswahrscheinlichkeit der subjektiven Erfolgswahrscheinlichkeit entspricht. Ist dies gegeben, hätte dem Modell zufolge das dispositionelle Leistungsmotiv für die motivationalen Tendenzen im CAT maximale Bedeutung, da im CAT üblicherweise eine individuelle objektive Erfolgswahrscheinlichkeit von 50 Prozent besteht. Das Leistungsmotiv hätte im CAT zudem bei den meisten Personen einen stärkeren Einfluss auf die motivationalen Tendenzen als im FIT. Personen mit hohem Erfolgsmotiv müssten im CAT eine höhere Motivation zeigen als im FIT, während Personen mit hohem Misserfolgsmotiv im CAT eine geringere Motivation zeigen müssten als im FIT. Weichen die objektive und die subjektive Erfolgswahrscheinlichkeit voneinander ab, stellt sich allerdings eine klare Ableitung von Annahmen aus dem Risikowahl-Modell in Bezug auf motivationale Unterschiede im FIT und im CAT als schwierig dar. Das Risikowahl-Modell liegt der Studie von Frey (2006) zu Motivation zur Testbearbeitung in CAT und FIT zugrunde, auf die in den Abschnitten 4.2 und 4.3 Bezug genommen wird.

Während die Erfolgstendenz eine aktivierende Kraft ist, wirkt die Misserfolgstendenz hemmend. Wegen der angenommenen Unabhängigkeit der beiden Motivkomponenten postuliert

das Risikowahl-Modell, dass in einer Leistungssituation immer beide motivationalen Tendenzen aktiviert werden und somit unausweichlich ein Annäherungs-Vermeidungs-Konflikt entsteht (vgl. Nähe zum Konfliktmodell nach Miller, 1944). Die Erfolgstendenz wird sozusagen stets um den Betrag der Misserfolgstendenz vermindert. Auch das Ergebnis dieses Konflikts umschreibt Atkinson formal, indem er die sogenannte *resultierende Tendenz* (RT) definiert als Summe der Erfolgs- und der Misserfolgstendenz: $RT = T_e+T_m = (M_e \times W_e \times A_e) + (M_m \times W_m \times A_m)$. Es sei angemerkt, dass die Misserfolgstendenz $T_m$ aufgrund der Definition von $A_m = -W_e$ stets negative Werte annimmt und es sich daher bei der Formel zur Berechnung von RT indirekt um eine Differenzbildung handelt. Die resultierende Tendenz bestimmt schließlich, ob eine Person sich leistungsbezogenen Aufgaben annähert (im Falle, dass die Erfolgstendenz größer ist als die Misserfolgstendenz und RT somit einen positiven Wert annimmt) oder ob sie sie vermeidet (im Falle, dass die Misserfolgstendenz gegenüber der Erfolgstendenz überwiegt und RT somit einen negativen Wert annimmt; Schmalt & Langens, 2009). Die resultierende Tendenz ist nach dem Risikowahl-Modell bei Aufgaben mittlerer subjektiver Schwierigkeit stets am stärksten ausgeprägt (vgl. Abbildung 3.2). Ist $T_e > T_m$, wird das Aufsuchen von Aufgaben mittlerer Erfolgswahrscheinlichkeit präferiert, ist $T_e < T_m$, würde ein Verlassen der Leistungssituation präferiert, da das Leistungsverhalten gehemmt wird. Ist das Verlassen der Leistungssituation nicht möglich, wie in Kompetenztestsituationen üblich, werden sehr einfache oder sehr schwierige Aufgaben relativ bevorzugt. Daraus, dass der Einfluss der Motivkomponente bei Aufgaben mittlerer wahrgenommener Schwierigkeit maximal ist, folgt, dass sich Unterschiede zwischen erfolgs- und misserfolgsmotivierten Personen insbesondere bei diesen Aufgaben in der resultierenden Motivation und im Leistungsverhalten niederschlagen; bei Aufgaben extremer wahrgenommener Schwierigkeiten lassen sich hingegen kaum motivbedingte Unterschiede feststellen.

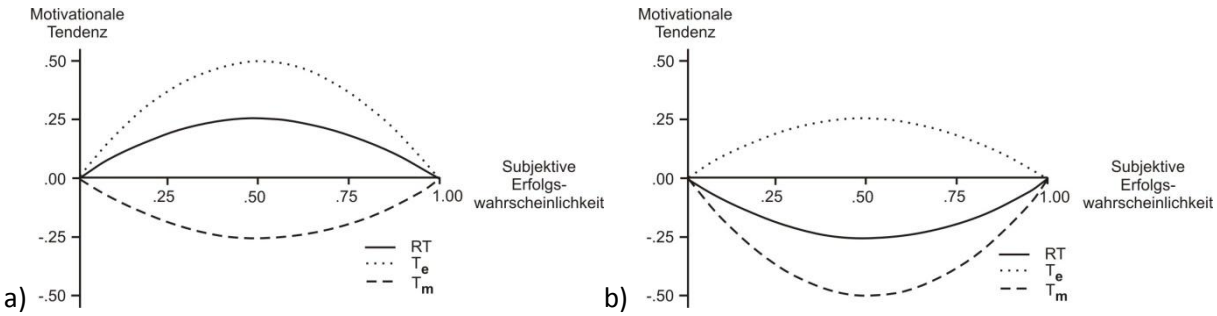



*Abbildung 3.2.* Erfolgstendenz ($T_e$), Misserfolgstendenz ($T_m$) und resultierende Tendenz (RT) von zwei hypothetischen Testpersonen mit überwiegender Erfolgstendenz ($T_e > T_m$; 3.2a) bzw. überwiegender Misserfolgstendenz ($T_m > T_e$; 3.2b) in Abhängigkeit von der subjektiven Erfolgswahrscheinlichkeit (nach Frey, 2006, S. 100; Schneider & Schmalt, 1994, S. 255).

Atkinsons Annahme, dass misserfolgsmotivierte Personen Aufgaben mittlerer wahrgenommener Schwierigkeit maximal meiden, regte eine kontroverse Diskussion an: Dem Schwierigkeitsgesetz der Motivation zufolge (Hillgruber, 1912; zitiert nach Rheinberg, 2008, S. 93) sollten Anstrengungsbereitschaft und Ausdauer (d. h. die „Intensitätsapekte“ der Leistungsmotivation; vgl. Rink, 1994) und die daraus resultierende Effizienz im Falle der unvermeidlichen Bearbeitung einer

Aufgabe umso höher sein, je größer die negative resultierende Tendenz ist (vgl. Anstrengungskalkulationsprinzip von Meyer, 1973). Denn eine Person zeigt dieser Annahme zufolge eine umso größere Anstrengungsabsicht, je schwieriger eine Aufgabe ist. Dies gilt jedoch nur bis zu einer Aufgabenschwierigkeit, von der die Person meint, sie mit maximaler Anstrengung gerade noch bewältigen zu können. Bei noch schwierigeren Aufgaben fällt die Anstrengungsbereitschaft rapide auf Null herab (Rink, 1994). Bezieht man diesen Ansatz auf erfolgsmotivierte Personen, folgt daraus, dass schwierige, aber nicht zu schwierige Aufgaben einen Anreiz für diese Personen haben, weil sie generell die Erfahrung gemacht haben, dass sie diese Aufgaben meistern können und daher nach einem positiveren affektiven Zustand streben. Einfache Aufgaben hingegen stellen keinen Anreiz dar, weil deren Bewältigung nicht außergewöhnlich ist und daher auch nicht zu der positiven Emotion Stolz führen kann (vgl. Brunstein, 2006). Während Atkinson diesem Ansatz zunächst selbst zustimmte (Atkinson, 1957), kehrte er sich später aus theoretischen Gründen davon ab und postulierte stattdessen, dass eine negative resultierende Tendenz auch die Anstrengungsbereitschaft und Ausdauer hemme (Atkinson & Feather, 1966). Empirisch hat sich jedoch das Anstrengungskalkulationsprinzip mehrfach bestätigt; auch die Misserfolgstendenz wirkt sich positiv auf das Aufgabenbearbeitungsverhalten aus, vermutlich weil Anstrengung eine Möglichkeit ist, den befürchteten Misserfolg und damit die antizipierte negative Emotion Scham zu vermeiden. Diese Befunde weisen darauf hin, dass sich die resultierende Tendenz nicht zwingend im tatsächlichen Verhalten in der Aufgabensituation widerspiegelt. Atkinson (1964) selbst erklärte scheinbar nicht theoriekonformes Verhalten bei Überwiegen der Misserfolgstendenz mit dem Wirksamwerden anderer, konfligierender Motive wie des Anschlussmotivs. Er ergänzte die Formel zur resultierenden Tendenz entsprechend um einen zusätzlichen additiven, allgemeinen Term, die extrinsische Tendenz ($T_{ex}$): $RT = T_e+T_m+T_{ex}$.

#### *3.2.1.4 Kritische Diskussion des Risikowahl-Modells*

Es ist ein Verdienst des Risikowahl-Modells, als erstes Erwartung-Wert-Modell dispositionelle und situationsspezifische Motivationskomponenten gleichwertig zu berücksichtigen. Mit der Konzeption der beiden Personenkomponenten Hoffnung auf Erfolg und Furcht vor Misserfolg als affektiv geprägte Merkmale und der vier Situationsvariablen Erfolgserwartung, Misserfolgserwartung, Erfolgsanreiz und Misserfolgsanreiz als eher kognitiv verankerte Merkmale nimmt Atkinson zudem eine deutliche Abkehr von der behavioristischen Sicht der frühen Leistungsmotivationstheorien vor.

Auch wenn das Risikowahl-Modell einen wichtigen Beitrag zur Leistungsmotivationsforschung geleistet hat und leistet, hat das Modell Schwächen. So reduziert die vollständige Determination der Wertkomponente durch die Erwartungskomponente den situationsspezifischen Teil des Modells auf ein einziges Konstrukt, nämlich die subjektive Erfolgswahrscheinlichkeit. Diesem Umstand ist wohl zuzuschreiben, dass die Wertkomponente des „Erwartung-Wert“-Modells als eigenständiges Konstrukt in etlichen Weiterentwicklungen und Folgestudien zur Leistungsmotivation stark vernachlässigt beziehungsweise kaum beachtet wurde (Brophy, 1999; Eccles & Wigfield, 1995; Middleton & Toluk, 1999; Schunk et al., 2008). Ferner widerspricht die inverse Verknüpfung von Erwartung und Wert zahlreichen empirischen Erkenntnissen, die einen positiven Zusammenhang zwischen diesen beiden Aspekten feststellen (Battle, 1966; Schunk et al., 2008; Wigfield & Eccles, 1992). Eccles und Wigfield (1995) kritisieren außerdem, dass Atkinson zumindest in seiner ursprünglichen Konzeption des Risikowahl-Modells 1957 die subjektive Erfolgserwartung per definitionem mit der objektiven Erfolgswahrscheinlichkeit gleichsetzt. Dies mag dazu geführt haben, dass in vielen theoretischen und empirischen Auseinandersetzungen mit dem Modell keine klare

Unterscheidung zwischen der subjektiven und der objektiven Erfolgswahrscheinlichkeit getroffen wurde und die Aussagekraft einiger Arbeiten daher hinterfragt werden muss.

Empirisch können die Ableitungen für erfolgsmotivierte Personen insgesamt recht gut bestätigt werden. Für misserfolgsmotivierte Personen erweist sich die empirische Befundlage hingegen als unklar, da diese Personen häufig keine eindeutige Präferenz für bestimmte Aufgabenschwierigkeiten zeigen. Brunstein und Heckhausen (2006) warnen allerdings vor einer vorschnellen Ablehnung des Modells für diese Personengruppe und äußern die Vermutung, dass hier ein Methodenproblem zugrunde liegen könnte, das in der Operationalisierung des Misserfolgsmotivs (respondente Erfassung) begründet ist. Auch dass viele Studien zur Prüfung des Risikowahl-Modells an studentischen Stichproben durchgeführt wurden, mag zu den heterogenen Ergebnissen bezüglich misserfolgsorientierter Personen geführt haben (Heckhausen, 1980). Empirisch zeigt sich, dass Personen im Allgemeinen Aufgaben mittlerer Schwierigkeit bevorzugen, auch wenn diese Präferenz bei erfolgsmotivierten Personen stärker ausgeprägt ist als bei misserfolgsmotivierten Personen (Meyer, Folkes & Weiner, 1976; Prochaska, 1998; Weiner, 1994). Eine Alternativerklärung für diesen Befund, die sich von Atkinsons Argument des Wirksamwerdens eines anderen Motivs abwendet und über die Annahmen des Risikowahl-Modells hinausgeht, formulieren Trope und Brickman (1975): Möglicherweise spielt weniger der antizipierte Affekt die entscheidende Rolle für die Präferenz für bestimmte Aufgaben, sondern vielmehr der erwartete Informationsgewinn (Diagnostizität) durch die Bearbeitung der Aufgabe (vgl. auch Meyer, 1973, 1984). Bei Aufgaben mittlerer Schwierigkeit erfährt man in der Regel am meisten über das eigene Können. Eine hohe Diagnostizität scheint dabei generell bevorzugt zu werden (Trope, 1975). Bei konstant gehaltener Diagnostizität werden nicht mittelschwierige, sondern einfache Aufgaben bevorzugt; für mittelschwierige und schwierige Aufgaben unterscheidet sich die Präferenz nicht (Trope, 1975). Aus diesen Ergebnissen schließt Weiner (1994), dass die Aufgabenwahl weniger affektiv als vielmehr kognitiv dominiert werde. Eine abschließende Klärung der Frage, ob sich die Aufgabenwahl eher an der Affektmaximierung oder am Informationsgewinn orientiert, steht noch aus (Brunstein & Heckhausen, 2006). Meyer et al. (1976) bestätigen hingegen die Bevorzugung von Aufgaben mittlerer subjektiver Erfolgswahrscheinlichkeit, die unabhängig von der objektiven Schwierigkeit der Aufgaben ist. Testpersonen präferieren im Allgemeinen Aufgaben, die sie selbst als für sie mittelschwierig einschätzen. In Abhängigkeit des Fähigkeitsselbstkonzeptes sind dies Aufgaben mit geringer bis hoher Schwierigkeit. Im Hinblick auf die Motivation zur Testbearbeitung in einem CAT erscheint es daher angezeigt, nicht nur die objektive Aufgabenschwierigkeit, sondern auch das Fähigkeitsselbstkonzept und die subjektive Erfolgswahrscheinlichkeit zu berücksichtigen. Dies geschieht in der vorliegenden Arbeit, indem der zentralen Fragestellung nachgegangen wird, wie sich CAT in Interaktion mit dem Fähigkeitsselbstkonzept auf die Motivation zur Testbearbeitung auswirkt (vgl. Abschnitt 5).

Streng genommen wurde das Risikowahl-Modell zwar lediglich für die prädezisionale Phase der Aufgabenwahl konzipiert, es wurde aber dennoch ebenso auf die Phase der Handlungsausführung angewendet. Für die Gültigkeit des Modells über den Entscheidungsmoment hinaus, und damit für das tatsächliche Leistungsverhalten, gibt es weder theoretische Begründungen noch empirische Belege (Beckmann & Heckhausen, 2006; Beckmann & Keller, 2009; Rink, 1994). Auch ist das Modell ursprünglich nur für den Fall konzipiert, dass keinerlei andere Motive außer dem Leistungsmotiv in der Leistungssituation angeregt werden und allein das Leistungsmotiv die Aufgabenwahl determiniert (Beckmann & Keller, 2009). Extrinsische Anreize oder nicht-leistungsthematische,

antizipierte Folgen der Aufgabenwahl bleiben außer Betracht. Dies sind unrealistische Annahmen, die auf die meisten Leistungssituationen nicht zutreffen dürften. Nicht plausibel erscheint zudem, dass misserfolgsmotivierte Personen dem Modell nach keinerlei Leistungsverhalten zeigen dürften. Wie Weiner (1994) anmerkt, widerspricht auch dies der Realität. Den letztgenannten beiden Kritikpunkten begegnete Atkinson zwar durch Erweiterung seines Modells um die extrinsische Tendenz als konfligierendes, alternativ wirksam werdendes Motiv. Die genaue Definition dieser Variable ließ er jedoch offen, was eine empirische Überprüfung dieses Modellteils unmöglich macht.

Insbesondere aufgrund der Schwierigkeit, das Aufgabenwahlverhalten bei misserfolgsmotivierten Personen anhand des Risikowahl-Modells zutreffend vorherzusagen, aber auch aufgrund der weiteren genannten Kritikpunkte an dem Modell, sind seit Atkinsons Publikation 1964 verschiedene Weiterentwicklungen entstanden (für einen Überblick siehe z. B. Weiner, 1994). Eine dieser Weiterentwicklungen ist das Erwartung-Wert-Modell der Leistungsmotivation (z. B. Eccles & Wigfield, 2002), das der vorliegenden Arbeit zugrunde liegt und das sich insbesondere zur Vorhersage schulischen Leistungsverhaltens bewährt hat (Brunstein & Heckhausen, 2006). Dieses Modell wird im folgenden Abschnitt erläutert.

### 3.2.2 Das Erwartung-Wert-Modell der Leistungsmotivation von Eccles & Wigfield (2002)

Dieser Abschnitt widmet sich dem Erwartung-Wert-Modell der Leistungsmotivation in der Version von Eccles und Wigfield (2002; vgl. auch ursprüngliches Modell von Eccles, Adler, Futterman, Goff, Kaczala et al., 1983). Dieses Modell lehnt sich an Atkinsons Risikowahl-Modell an, indem Merkmale leistungsmotivierten Verhaltens durch das Zusammenwirken von Erwartungs- und Wertvariablen erklärt werden (Schunk et al., 2008). Dabei beschränkt sich das Modell jedoch nicht auf die Prognose von Wahlverhalten, sondern es schließt die Vorhersage von Testleistung explizit mit ein. Außerdem ist das Modell gegenüber Atkinsons Ansatz stärker situativ ausgerichtet, auch wenn der dispositionellen Personenkomponente weiterhin eine wichtige Bedeutung beigemessen wird (Möller, 2008; Möller & Schiefele, 2004; Schunk et al., 2008). Ein weiterer Unterschied zum Risikowahl-Modell ist, dass die Wertkomponente ausdrücklich als eigenständiges Konstrukt statt nur als Komplement zur Erwartungskomponente beachtet wird. Zudem wird ein positiver statt negativer Zusammenhang zwischen beiden Komponenten postuliert (Wigfield, 1994).

Das Erwartung-Wert-Modell der Leistungsmotivation ist breit angelegt, indem es das soziokulturelle Umfeld und die Entwicklungsgeschichte eines Individuums zur Prognose und Erklärung leistungsmotivierten Verhaltens einbezieht. Es wurde ursprünglich dafür entwickelt, Leistung und Wahlverhalten von adoleszenten Personen im Bereich mathematischer Kompetenz zu verstehen (Wigfield, 1994; Wigfield & Eccles, 1992; Wigfield & Eccles, 2002). Demzufolge nimmt das Modell deutlichen theoretischen Bezug auf schulisches Leistungsverhalten und hat sich auch empirisch für die Anwendung auf den schulischen Leistungskontext als sehr fruchtbar erwiesen (Brunstein & Heckhausen, 2006).

#### *3.2.2.1 Darstellung des Gesamtmodells*

Die wichtigsten Begriffe des Erwartung-Wert-Modells der Leistungsmotivation werden im Folgenden definiert. Anschließend wird auf die postulierten Zusammenhänge im Modell eingegangen.

Das vollständige Erwartung-Wert-Modell der Leistungsmotivation nach Eccles und Wigfield (2002) ist Abbildung 3.3 dargestellt. Die Erfolgserwartung ist eine verhaltenswirksame, subjektive Einschätzung der Wahrscheinlichkeit, eine Aufgabe erfolgreich zu bearbeiten. Dabei ist der Begriff „Aufgabe“ weit gefasst zu verstehen: Er kann sich auf eine einzelne, spezifische Aufgabe beziehen, er kann jedoch auch einen Test in einer bestimmten Domäne meinen (Schunk et al., 2008; Wigfield & Eccles, 2002). Diese subjektive Einschätzung muss nicht mit der objektiven Erfolgswahrscheinlichkeit übereinstimmen. Die Genauigkeit der Einschätzung kann einen Effekt auf die nachfolgende Leistung haben. So scheint sich eine relativ akkurate, jedoch leicht ins positive verzerrte Einschätzung der eigenen Erfolgswahrscheinlichkeit förderlich auf die Leistung auszuwirken, während eine deutliche Über- oder Unterschätzung die Leistung beeinträchtigen kann (Schunk et al., 2008; vgl. Abschnitt 3.2.2.4). Die Erwartungskomponente steht bei Eccles und Wigfield theoretisch in enger Beziehung zu Konzepten von Kompetenzüberzeugungen, wie zum Fähigkeitsselbstkonzept oder zur Selbstwirksamkeitserwartung (in dem Modell unter Selbstschemata subsumiert, vgl. Abbildung 3.3) und lässt sich empirisch oft nur schwer von ihnen trennen. Das *Fähigkeitsselbstkonzept* ist definiert als wahrgenommene eigene Leistungsausprägung in einer bestimmten Domäne (Marsh, 1993). Es beinhaltet theoretisch sowohl kognitive als auch affektive Aspekte, wobei empirisch häufig nur der kognitive Aspekt des Selbstkonzepts operationalisiert wird (vgl. Bong & Clark, 1999; Möller & Trautwein, 2009). Das Fähigkeitsselbstkonzept bildet sich im Laufe der Zeit durch Erfahrungen des Selbsts mit der Umwelt und durch die Interpretation dieser Erfahrungen heraus. Seine Ausprägung ist stark von dem jeweiligen Bezugsrahmen abhängig. So wird das Fähigkeitsselbstkonzept von Schülerinnen und Schülern maßgeblich durch den sozialen Vergleich der eigenen Leistung mit der Leistung der Mitschülerinnen und Mitschüler derselben Klasse oder Schulart beeinflusst (vgl. Köller, 2004; Möller & Schiefele, 2004). Die *Selbstwirksamkeitserwartung* hingegen ist als rein kognitives Merkmal konzipiert und entspricht der individuellen Überzeugung, spezifische Tätigkeiten oder Aufgaben aufgrund eigener Fähigkeiten erfolgreich ausführen zu können (Bandura, 1997). Sie ist relativ unabhängig von sozialen und dimensionalen Vergleichsprozessen und wird eher von Erfahrungen mit ähnlichen Aufgaben oder Tätigkeiten und von absoluten Leistungsmaßstäben beeinflusst (Bong & Clark, 1999; Möller, Pohlmann, Köller & Marsh, 2009; Möller & Schiefele, 2004).

Wigfield und Eccles (2000) weisen darauf hin, dass ihr Konzept der Erfolgserwartung eine größere konzeptionelle Nähe zum Fähigkeitsselbstkonzept hat als zur Selbstwirksamkeitserwartung, da sie das Merkmal eher als domänenspezifisch statt als tätigkeitsspezifisch verstanden wissen wollen. Erfahrungen mit bestimmten Aufgabenarten bestimmen, wie die eigene Leistungsausprägung in dem jeweiligen Aufgabengebiet eingeschätzt wird. Aus dieser Einschätzung entsteht die Erwartungshaltung, die die Person zukünftigen Aufgaben desselben Inhaltsgebiets gegenüber hat. Dies geht konform mit dem Hinweis Brandstätters (2009), dass sich die (tätigkeitsspezifische) Selbstwirksamkeitserwartung eher auf die Lernmotivation und Lernziele auswirkt, während das Fähigkeitsselbstkonzept der bedeutsamere Prädiktor für Leistungsmotivation und Leistungsziele ist (vgl. auch Heckhausen & Heckhausen, 2006; ein gegenteiliges Fazit ziehen jedoch Bong & Clark, 1999, aus einer Literaturzusammenschau).

Fähigkeitsselbstkonzept und Erfolgserwartung sind zwei theoretisch-konzeptionell unterschiedliche Merkmale: Das Fähigkeitsselbstkonzept hat einen etwas breiteren Geltungsbereich und fokussiert stärker auf die gegenwärtige Leistung, während sich die Erfolgserwartung stärker auf direkt bevorstehende Aufgaben und die zukünftige Leistung bezieht (Wigfield, 1994; Wigfield &

Eccles, 2000; vgl. auch Bong & Clark, 1999). Diese Unterscheidung ist jedoch empirisch oft nicht zu finden (z. B. Eccles & Wigfield, 1995). Testpersonen scheinen beispielsweise keinen Unterschied in der Einschätzung ihrer mathematischen Fähigkeiten und der Erwartung des Abschneidens in einem Mathematiktest zu machen (Schunk et al., 2008). Eine Längsschnittstudie von Meece, Wigfield und Eccles (1990) zeigt, dass sich die Erfolgserwartung in Mathematik durch das vorjährig gemessene mathematische Fähigkeitsselbstkonzept von Schülerinnen und Schülern vorhersagen lässt.

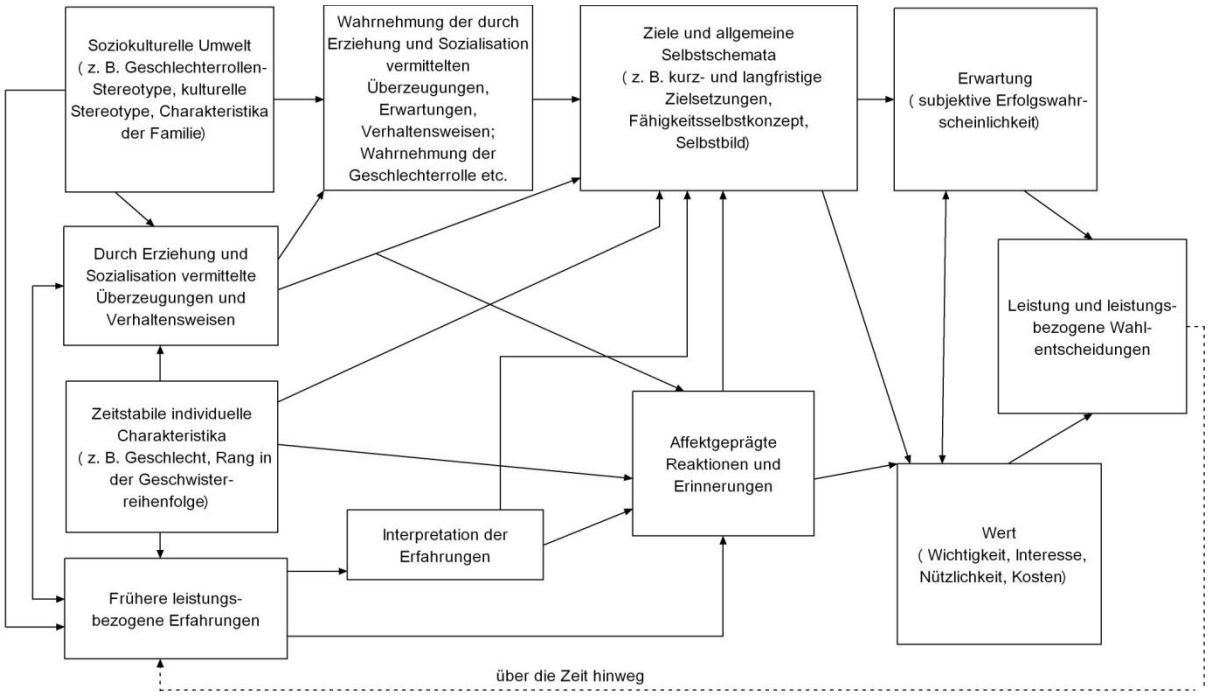



*Abbildung 3.3*. Erwartung-Wert-Modell der Leistungsmotivation (nach Eccles & Wigfield, 2002, S. 119; eigene Übersetzung).

Theoretisch und empirisch klar von den Erfolgserwartungen abgrenzbar ist die *Wertkomponente* des Modells, die sich auf die Überzeugungen einer Person bezieht, warum es sich lohnen könnte, eine Aufgabe zu bearbeiten beziehungsweise Leistung zu zeigen. Die Wertkomponente ist insbesondere für das Aufgabenwahlverhalten, also für leistungsbezogene Entscheidungen, relevant (Wigfield & Eccles, 2002). Bei der Wertkomponente unterscheiden Eccles und Wigfield vier verschiedene Aspekte: Wichtigkeit, Interesse und Nützlichkeit sowie Kosten. Die Autoren betonen, dass die einzelnen Wertaspekte vor unterschiedlichen theoretischen Hintergründen entstanden sind (Wigfield & Eccles, 2000). Hinsichtlich ihrer empirischen Unterscheidbarkeit gibt es jedoch divergierende Befunde. So gibt es für die drei erstgenannten Hinweise, dass sie sich empirisch trennen lassen (Eccles & Wigfield, 1995; Schunk et al., 2008). Es gibt jedoch auch Studien, in denen zumindest Wichtigkeit und Interesse faktorenanalytisch nicht unterschieden werden können (z. B. Köller, Daniels, Schnabel & Baumert, 2000). Der Kosten-Aspekt wird in empirischen Studien kaum berücksichtigt (Wigfield & Eccles, 2000).

*Wichtigkeit* ist definiert als die subjektive Bedeutsamkeit, eine Aufgabe erfolgreich zu bearbeiten. Sie wird unter anderem durch das Fähigkeitsselbstkonzept beeinflusst: Wenn ich glaube, dass ich gut in Mathematik bin, ist es mir (im Sinne eines hedonistischen Ansatzes; vgl. Kuhl, 1983; Prochaska, 1998; Schmalt & Langens, 2009) wichtig, in einer Mathematikaufgabe gut abzuschneiden. Diese Aufgabe gibt der Person die Möglichkeit, für sie wichtige Aspekte ihres eigenen Ichs, ihres Selbsts, auszudrücken oder zu bestätigen (Eccles, 2005). Dieser Zusammenhang kann auch stereotypische Einstellungen oder Bewertungen erklären, beispielsweise weshalb Mädchen, die ein feminines Rollen- beziehungsweise Selbstkonzept haben, Physik im Allgemeinen wenig wertschätzen (z. B. Kessels, Rau & Hannover, 2006). Auch das dispositionelle Leistungsmotiv (Hoffnung auf Erfolg bzw. Furcht vor Misserfolg; vgl. Abschnitt 3.2.1) beeinflusst die Ausprägung der Wichtigkeit. Dieser Wertaspekt kommt inhaltlich am ehesten Atkinsons Verständnis der Wertkomponente nahe (vgl. Abschnitt 3.2.1).

Der zweite Wertaspekt, *Interesse*, bezieht sich auf die Freude, die eine Person verspürt, wenn sie eine Aufgabe bearbeitet, auf die antizipierte Freude des Bearbeitens (Eccles, 2005) oder auf das genuine Interesse an dem Inhalt einer Aufgabe. Interesse steht damit in engem Zusammenhang mit dem Inhalt und der eigentlichen Ausführung einer Aufgabe und dem damit verbundenen Affekt, weniger aber mit den antizipierten Konsequenzen des Leistungsergebnisses. In hoher Intensität ist Interesse vergleichbar mit dem Flow-Erleben (vgl. Csikszentmihalyi, 1990). Interesse als Wertaspekt ähnelt der tätigkeits- und gegenstandsbezogenen intrinsischen Motivation (Schiefele, 1996), es ist jedoch konzeptuell nicht mit ihr identisch (Eccles, 2005). Im Modell von Eccles und Wigfield (2002) steht der affektive Aspekt im Vordergrund, während in der Interessekonzeption von Schiefele (1996) der Aspekt des inhaltlichen, gegenstandsbezogenen Interesses in Verbindung mit der Wichtigkeit des Gegenstands dominiert (vgl. auch Köller et al., 2000).

*Nützlichkeit*, der dritte Wertaspekt, wird verstanden als Maß dafür, wie sehr die erfolgreiche Bearbeitung einer Aufgabe zum Erreichen persönlicher zukünftiger Ziele oder Karriereambitionen beiträgt. Dieser Aspekt hat im Gegensatz zum Interesse keinen intrinsischen, sondern einen extrinsischen Fokus. Die Nützlichkeit bezieht sich auf das Ergebnis der leistungsmotivierten Handlung und wird durch das Vorhandensein persönlicher Konsequenzen des Leistungshandelns angeregt.

Neben diesen drei positiv formulierten Wertaspekten betrachten Wigfield und Eccles noch die *Kosten*, die jemand investieren muss, um eine Aufgabe erfolgreich zu bearbeiten. Dieses Konstrukt bezieht sich auf wahrgenommene negative Aspekte der Aufgabenbearbeitung, wie den Verzicht auf alternative Tätigkeiten, die zu investierende Anstrengung oder mögliche antizipierte negative Gefühle (Testängstlichkeit, Scham beim Versagen; siehe konzeptuelle Nähe zu Furcht vor Misserfolg).

Die Autoren gehen davon aus, dass alle vier Wertaspekte gleichzeitig wirken und somit den Wert einer Aufgabenbearbeitung bestimmen. Dabei korrelieren Wichtigkeit, Interesse und Nützlichkeit mäßig bis hoch positiv miteinander ($.56 < r < .79$), wohingegen sie gering bis mäßig negativ mit den Kosten zusammenhängen ($-.32 < r < -.13$; Eccles & Wigfield, 1995). Innerhalb der Wertkomponente zeigt das Interesse die größte Situationsspezifität und die stärkste Verbindung zur Erwartungskomponente (Wigfield & Eccles, 2002).

Auf der rechten Seite von Abbildung 3.3 sind die eben beschriebenen zentralen Komponenten der aktuellen Leistungsmotivation, Erwartung und Wert, dargestellt, die das Leistungsverhalten und

leistungsbezogene Wahlentscheidungen vorhersagen. Die übrigen Elemente des Modells illustrieren die Entstehung von Erwartung und Wert durch ein Zusammenspiel von intern-subjektiven und extern-soziokulturellen Einflussgrößen: Unter *Zielen und allgemeinen Selbstschemata* sind neben den oben angesprochenen individuellen Kompetenzüberzeugungen (Fähigkeitsselbstkonzept, Selbstwirksamkeitserwartung) kurz- und langfristige Zielsetzungen sowie das generelle leistungsbezogene Selbstbild einer Person subsumiert. Die Ziele entsprechen kognitiven Repräsentationen dessen, was eine Person anstrebt (vgl. Abschnitt 3.1). Ein kurzfristiges Ziel bezieht sich dabei auf proximale Ereignisse, wie einen bevorstehenden Test möglichst gut zu bestehen. Distale Ziele sind beispielsweise Vorstellungen über den späteren Beruf. Schunk et al. (2008) betonen, dass die hier angesprochenen Ziele nicht mit Zielorientierungen gleichzusetzen sind. Das Selbstbild beinhaltet eine so genannte Idealvorstellung einer Person von sich selbst, bezogen auf die eigene Persönlichkeit und Identität (Schunk et al., 2008): Wer bin ich und wie möchte und kann ich sein?

*Affektgeprägte Reaktionen und Erinnerungen* einer Person wirken vor allem auf die Wertkomponente, aber auch auf die Ziele und Kompetenzüberzeugungen ein. Sie ergeben sich aus den subjektiven Interpretationen früher gemachter Erfahrungen in ähnlichen Leistungssituationen, bilden sich über die Zeit hinweg und sind relativ stabil. Sie ähneln damit der von Atkinson (1964) vorgenommenen Konzeption der Leistungsmotivkomponenten Hoffnung auf Erfolg und Furcht vor Misserfolg als antizipiertem positivem oder negativem Affekt (vgl. auch Wigfield & Eccles, 1992). Im Zuge der Interpretation früherer Erfahrungen spielen auch Attributionen, das heißt, subjektive Ursachenzuschreibungen, eine wichtige Rolle (vgl. Attributionstheorie von Weiner, 1994).

Die Ziele und die allgemeinen Selbstschemata gehen unter anderem auf die subjektive Wahrnehmung der in Erziehung und Sozialisation vermittelten Überzeugungen, Erwartungen und Verhaltensweisen einer Person zurück: Nimmt eine Schülerin beispielsweise wahr, dass ihr Physiklehrer Mädchen im Unterricht weniger zutraut als Jungen (unabhängig davon, ob dies der tatsächlichen Auffassung und dem Verhalten des Physiklehrers entspricht), kann sich dies in ihrem Selbstbild niederschlagen. Das Modell berücksichtigt schließlich noch die soziokulturelle Umwelt und zeitstabile Charakteristika einer Person wie den familiären Hintergrund, das Geschlecht, das soziale Umfeld oder die Einstellung des sozialen Umfelds zu Leistung. Insgesamt wird der Leistungsmotivationsprozess als sich stetig wiederholender Kreislauf aus manifesten, objektiven Umständen und Ereignissen und deren Interpretation beziehungsweise Wahrnehmung durch die Person dargestellt (Abbildung 3.3; für eine detailliertere Beschreibung des Gesamtmodells siehe Eccles, 1993).

Hinsichtlich der Beziehungen zwischen den einzelnen Elementen des Modells ist zusammenzufassen, dass die Erwartung einer Person, erfolgreich zu sein, und der Wert, den sie einem Erfolg beimisst, wesentliche und direkte Bestimmungsstücke der Motivation zur Aufgabenbearbeitung und der Aufgabenwahl und damit des Leistungsverhaltens sind (Brunstein & Heckhausen, 2006; Wigfield & Eccles, 2002). Die Erwartungen und Werte wiederum entstehen durch ein komplexes Zusammenspiel aus persönlichen und soziokulturellen Einflussfaktoren. Erziehungs- und Kultureinflüsse wirken demnach indirekt über die Erwartungs- und Wertkomponenten auf das Leistungsverhalten ein. Der rekursive Pfeil von der Leistung und leistungsbezogenen Wahlentscheidungen zu früher gemachten Erfahrungen verdeutlicht darüber hinaus die dynamische Konzeption des Modells (Eccles, 2005), die den dynamischen Aspekt der Definition von Leistungsmotivation abbilden kann (vgl. Abschnitt 3.1) und Kuhls (1983) Forderung aufgreift, nicht

nur prä-intentionale, sondern auch post-aktionale Aspekte der Leistungsmotivation zu berücksichtigen.

#### *3.2.2.2 Empirische Befunde*

Das Erwartung-Wert-Modell der Leistungsmotivation konnte anhand verschiedener Merkmale des schulischen Leistungsverhaltens erfolgreich validiert werden (Wigfield & Eccles, 2000): Bereits zu Beginn der Grundschulzeit unterscheiden Schulkinder auf domänenspezifischer Ebene deutlich zwischen der Erwartung, ob sie in einer Aufgabe „gut" sind, und den Werten, wie sehr sie den Erfolg in einer Aufgabe schätzen (Wigfield, 1994; Wigfield & Eccles, 2002). Selbst wenn die Ausgangsleistung von Schülerinnen und Schülern kontrolliert wird, sagen subjektive Kompetenzüberzeugungen, Erfolgserwartungen und Werte zukünftige Schulleistungen beziehungsweise die Präferenz für bestimmte Schulfächer vorher. Beispielsweise bilden die subjektiv wahrgenommene Kompetenz und die Erfolgserwartung die stärksten Prädiktoren für spätere Zeugnisnoten (z. B. in Mathematik oder Englisch; z. B. Eccles & Wigfield, 2002; Schunk et al., 2008; Wigfield & Eccles, 2000) und übertreffen sogar vorherige Noten in ihrer Vorhersagekraft. Diese Ergebnisse wurden im Klassen- beziehungsweise Schulsetting beobachtet und konnten in Längsschnittstudien mehrfach bestätigt werden (Wigfield & Eccles, 2000). Es ist daher eine hohe ökologische Validität und Generalisierbarkeit dieser Befunde anzunehmen. Dabei ist von einem reziproken Zusammenhang auszugehen, da sich zum Beispiel Selbstkonzept und Leistung gegenseitig positiv beeinflussen (vgl. rekursiver Pfeil in Abbildung 3.3; Guay, Marsh & Boivin, 2003; Schunk et al., 2008). Auch für den positiven Einfluss von Erwartungskomponente und Selbstkonzept auf Testleistungen in standardisierten Leistungstests gibt es vielfältige empirische Belege (Eccles, 2005; vgl. Schunk et al., 2008; Valentine, DuBois & Cooper, 2004; Wigfield & Eccles, 1992; vgl. Abschnitt 4). Man kann davon ausgehen, dass Erfolgserwartung und subjektive Kompetenzüberzeugungen wichtige Mediatoren zwischen soziokulturellem Umfeld und aktuellem Leistungsverhalten darstellen (vgl. Abbildung 3.3; Schunk et al., 2008).

Die Wertkomponente beeinflusst die Ausdauer bei begonnener Aufgabenbearbeitung (Schunk et al., 2008; Wigfield & Eccles, 2000) und erweist sich als bedeutsamer Prädiktor von Kurswahl- oder Aufgabenwahlentscheidungen (z. B. in Mathematik, Physik, Englisch, Sport: Eccles & Wigfield, 2002). Diesbezüglich spielt die Erwartung im Vergleich häufig eine geringere Rolle (Schunk et al., 2008). Das Fähigkeitsselbstkonzept allerdings, das empirisch eng mit der Erwartungskomponente assoziiert ist, hat sich als sehr bedeutsam für die Vorhersage von Entscheidungen erwiesen. Dies wurde beispielsweise anhand von Kurswahlentscheidungen in der schulischen Oberstufe gezeigt (Köller et al., 2000).

Hinsichtlich der Entwicklung der Erwartungs- und der Wertkomponente der Leistungsmotivation über die Zeit zeigen Wigfield, Eccles und Kollegen in Quer- und Längsschnittstudien, dass die absolute Höhe der Kompetenzüberzeugung (z. B. in Mathematik) im Verlauf der Schulzeit abnimmt; für die Selbstwirksamkeitserwartung hingegen finden sie eine gegenläufige Entwicklung (z. B. Eccles, Wigfield, Harold & Blumenfeld, 1993; Wigfield, 1994; Wigfield & Eccles, 2000, 2002). In Bezug auf die Wertkomponente zeigt sich ein differenziertes Bild: Während Wichtigkeit und Nützlichkeit von Mathematik in der Sekundarstufe geringer wertgeschätzt werden als in der Primarstufe, verändert sich das Interesse an Mathematik nicht wesentlich. Generell ist jedoch ein Trend zu erkennen, dass sowohl Erwartung als auch Wert in Bezug auf viele

domänenspezifische Leistungen im Verlauf der Schulzeit eher ab- als zunehmen (Wigfield, 1994). Besonders deutlich zeichnet sich dies in der Adoleszenz ab, also in der Sekundarstufe (Wigfield & Eccles, 2000). Dies lässt es insbesondere bei empirischen Bildungsforschungsstudien, die Jugendliche untersuchen (wie z. B. PISA), als ratsam erscheinen, motivationale Effekte bei der Interpretation der Testleistung zu berücksichtigen. Brunstein und Heckhausen (2006) sowie Schunk et al. (2008) weisen darauf hin, dass sich in dem negativen Entwicklungstrend in Bezug auf die Erwartungskomponente und die subjektiven Kompetenzüberzeugungen eine Anpassung des Selbstbilds an die Realität abbildet: Während jüngere Kinder generell zu einer deutlichen Überschätzung ihrer Fähigkeiten neigen, gleicht sich diese subjektive Einschätzung im Verlauf der Schulzeit stärker den objektiven Gegebenheiten an (vgl. auch Möller & Trautwein, 2009). Zu dieser Interpretation passt, dass die Korrelation zwischen Erwartung beziehungsweise Selbstkonzept und Leistung in der Sekundarstufe deutlich höher ist als in der Primarstufe. Bei Jugendlichen spielen motivationale Aspekte demnach für die aktuelle Leistung eine wichtigere Rolle als bei Grundschulkindern (vgl. Wigfield, 1994). Brunstein und Heckhausen (2006) meinen, dass die Verringerung der Erwartung und des Fähigkeitsselbstkonzepts aus pädagogischer Sicht dann als Warnzeichen wahrgenommen werden sollte, wenn Jugendliche sich unnötigerweise unterschätzen und glauben, den schulischen Anforderungen mit ihren eigenen Fähigkeiten nicht mehr gerecht werden zu können.

Zur prognostischen Bedeutsamkeit der Erwartungs- und der Wertkomponente des Modells ist zusammenzufassen, dass die Erwartungskomponente eher für aktuelle Leistung und kognitive Anstrengung wichtig ist, während die Wertkomponente vor allem Wahlentscheidungen und die Ausdauer bei begonnener Aufgabenbearbeitung beeinflusst (Schunk et al., 2008). Während zunächst – wie im Risikowahl-Modell (Atkinson, 1964; vgl. Abschnitt 3.2.1) – explizit von einer inversen Beziehung zwischen Erwartung und Wert ausgegangen wurde, mehren sich die empirischen Belege, dass die beiden Komponenten positiv miteinander korrelieren (Wigfield et al., 1997) oder dass sich in Domänen wie Lesen oder Sport möglicherweise die Erwartung positiv auf den Wert auswirkt (Eccles & Wigfield, 1995; Schunk et al., 2008; Wigfield & Eccles, 2002; vgl. auch Möller & Schiefele, 2004). Es erscheint plausibel, dass Kinder und Jugendliche leistungsbezogene Tätigkeiten im Verlaufe ihrer Entwicklung mehr zu schätzen lernen, in denen sie gut sind. Denn zum einen werden diese Tätigkeiten bei Erfolg mit positivem Affekt assoziiert, was eine Art klassischen Konditionierungsprozess in Gang setzen kann (Eccles et al., 1983). Zum anderen kann eine „Ab-Wertung“ schwieriger Tätigkeiten, die mit individuellem Misserfolg konnotiert sind, zur Aufrechterhaltung eines positiven Selbstbildes beitragen (Eccles & Wigfield, 2002). Insbesondere in der Adoleszenz spielen Erwartung und Wert als Komponenten der Leistungsmotivation eine bedeutsame Rolle für das Leistungsverhalten. Generell gilt das Erwartung-Wert-Modell nach Eccles und Wigfield als empirisch gut validiert. Eccles (2005) vermutet, dass es aufgrund seiner dynamischen Konzeption nicht nur auf Entwicklungen über die Lebensspanne, sondern auch auf den Prozess der Aufgabenbearbeitung in einem Kompetenztest und auf hochfrequente Motivationsmessungen („moment-to-moment“) anwendbar ist (vgl. auch Abschnitt 3.2.2.4).

#### *3.2.2.3 Fazit zum Gesamtmodell*

Zusammenfassend betrachtet greifen Eccles und Wigfield (2002) die wesentlichen Komponenten des Risikowahl-Modells von Atkinson auf, indem sie Erfolgserwartungen und Werte als zentrale Prädiktoren des Leistungsverhaltens definieren. Sie erweitern das bei Atkinson rein kognitiv-affektive Modell jedoch um eine soziale Perspektive und werden damit der gegenwärtig gültigen Auffassung von Motivation als sozial-kognitiv-affektive Größe gerecht (Schunk et al., 2008). Auch stellen sie

gegenüber Atkinson ein elaborierteres Konzept der Erwartung- und Wertvariablen vor. Diese Variablen entwickeln sich durch kognitive Prozesse (Attributionen früheren Leistungsverhaltens, Wahrnehmungen der sozialen Umwelt; Schunk et al., 2008), welche wiederum im Rahmen des sozialen Umfelds einer Person entstehen. Die Motivkomponente aus Atkinsons Modell findet sich bei Eccles und Wigfield in den affektiven Erinnerungen und Erfahrungen wieder. Erwartung und Wert sowie die ihnen vorauswirkenden Leistungsziele, Kompetenzüberzeugungen, Wahrnehmungen und Interpretationen der Umwelt und der eigenen Erfahrungen sind – im Gegensatz zum manifesten Leistungsverhalten – als interne, nicht direkt beobachtbare kognitive Prozesse zu verstehen. Die Aspekte des soziokulturellen Umfelds und der Erziehung sowie frühere Erlebnisse entsprechen wiederum externen, beobachtbaren Einflüssen (Schunk et al., 2008).

Insgesamt geht das hier dargestellte Erwartung-Wert-Modell in seiner Komplexität und Aktualität deutlich über das Risikowahl-Modell hinaus. Das besondere Verdienst der Arbeit von Eccles und Wigfield ist, dass sie Erwartung und Wert als zwei eigenständige, positiv miteinander korrelierte Einflussgrößen der Leistungsmotivation definiert haben und die Domänenspezifität dieser Konstrukte angemessen im Modell berücksichtigen (Brunstein & Heckhausen, 2006). Dies kann beispielsweise erklären, weshalb die Leistungsmotivation einer Schülerin im Fach Deutsch anders ausgeprägt sein mag als im Fach Physik. Darüber hinaus verdeutlicht das Modell, dass es notwendig ist, neben der klassischen Situationsvariable „Aufgabenschwierigkeit“ weitere erwartungs- und anreizrelevante Größen in die Analyse der Leistungsmotivation mit einzubeziehen (Brunstein & Heckhausen, 2006; Eccles & Wigfield, 1995). Ferner ist das Modell dynamisch konzipiert und bildet dadurch gut ab, dass Motivation als fortwährender, zirkulärer Prozess aus Antizipation und Evaluation zu verstehen ist. Eine Person passt sich ständig aufs Neue und in Abhängigkeit von ihrer dispositionellen Motivation an die aktuelle Situation an (Middleton & Toluk, 1999). Dies entspricht dem Gedanken des Grundmodells motivierten Verhaltens in einer Testsituation (Abbildung 3.1), Motivation als Ergebnis der Interaktion von Personen- und Situationsmerkmalen zu betrachten. Zusammengefasst bietet das Erwartung-Wert-Modell der Leistungsmotivation einen soliden theoretischen Ausgangspunkt, um motivationale Prozesse in adaptiven und nicht-adaptiven Leistungstests zu untersuchen.

#### *3.2.2.4 Eingrenzung des Gesamtmodells auf die Motivation zur Testbearbeitung*

Wie aus dem vorherigen Abschnitt deutlich wird (vgl. Abbildung 3.3), ist das Erwartung-Wert-Modell der Leistungsmotivation von Eccles und Wigfield (2002) sehr komplex. Denn es berücksichtigt nicht nur die kognitiv-affektiven Prozesse, die innerhalb einer Person in einer Leistungssituation ablaufen und zum Leistungsverhalten führen, sondern es bezieht auch soziokulturelle Umfeldfaktoren mit ein. Eine empirische Gesamtbetrachtung des Modells würde den Rahmen der vorliegenden Arbeit übersteigen. Außerdem sind einige Aspekte des Gesamtmodells für ein Verständnis der bei einer Testung ablaufenden motivationalen Prozesse nicht nötig. Im Folgenden wird deshalb auf einen Teil des Modells fokussiert. Dies kann in der Leistungsmotivationsforschung als üblich angesehen werden (vgl. z. B. Eccles, 2005; Eccles & Wigfield, 2000; Middleton & Toluk, 1999; Möller & Schiefele, 2004). Dieses Teilmodell ist nicht als geschlossenes System zu verstehen; selbstverständlich werden die einzelnen Elemente des Teilmodells über die hier dargestellten, testsituationsspezifischen Zusammenhänge hinaus auch von den im Gesamtmodell berücksichtigten Faktoren beeinflusst. Das Teilmodell wird nachstehend dargestellt, und die einzelnen Konstrukte werden bezüglich der Vorhersage der Motivation zur Testbearbeitung spezifiziert.

Im Mittelpunkt des Modells von Eccles und Wigfield stehen die Erwartungs- und die Wertkomponente, die den Kern der aktuellen Leistungsmotivation bilden. Es ist davon auszugehen, dass diese beiden Komponenten auch in einer Leistungsteststituation die zentralen Aspekte der Motivation zur Testbearbeitung sind (Abbildung 3.4).

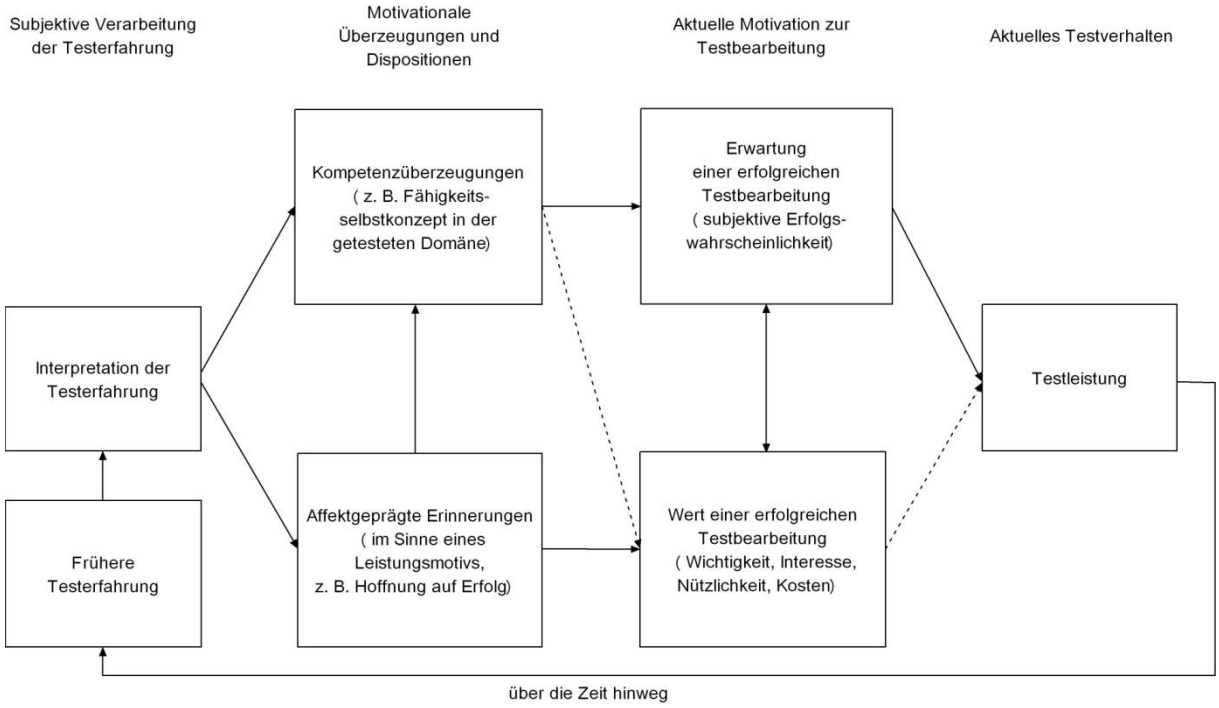



*Abbildung 3.4*. Teilmodell des Erwartung-Wert-Modells (Eccles & Wigfield, 2002) zur Vorhersage der Motivation zur Testbearbeitung.

Die Erwartung drückt aus, für wie wahrscheinlich es eine Testperson hält, in einem Leistungstest erfolgreich abzuschneiden.[2] Mit dem Wert wird beschrieben, wie wichtig, wie interessant und wie nützlich die erfolgreiche Bearbeitung des Tests für die Person ist. Auch die Kosten für eine erfolgreiche Testbearbeitung, beispielsweise in Form der aufzuwendenden Anstrengung, können die Wertkomponente prägen. Der Einfluss der Erfolgserwartung auf die Leistung ist deutlich stärker als der des Werts, was in Abbildung 3.4 durch die gestrichelte Linie zwischen Wert und Testleistung gekennzeichnet ist.

Die aktuelle Motivation zur Testbearbeitung wird durch motivationale Überzeugungen und Dispositionen beeinflusst. So hängt die Erfolgserwartung von den individuellen Kompetenzüberzeugungen ab, die sich beispielsweise im domänenspezifischen Fähigkeitsselbstkonzept der Testperson ausdrücken. Auch die Selbstwirksamkeitserwartung beeinflusst die Erwartung, in einem Test erfolgreich zu sein (vgl. Möller & Schiefele, 2004), doch sehen Wigfield und Eccles (2000) eine größere theoretische Nähe zwischen Selbstkonzept und

[2] Da mit dem Teilmodell die Motivation zur Testbearbeitung erklärt werden soll, stellt der Test die Bezugsgröße für die jeweiligen Definitionen dar (vgl. Abschnitt 3.2.2.1; Wigfield & Eccles, 2000).

Erwartung. Diese Kompetenzüberzeugungen wirken sich möglicherweise, in weit schwächerem Ausmaß, auch auf die Wertkomponente aus. Deutlich stärker wird der Wert jedoch durch die affektgeprägten Erinnerungen bestimmt, die im Sinne eines relativ zeitstabilen Leistungsmotivs (z. B. Hoffnung auf Erfolg) zu interpretieren sind. Die motivationalen Überzeugungen und Dispositionen werden durch die subjektive Verarbeitung früherer Testerfahrungen in ähnlichen Testsituationen geprägt. Während der Schulzeit entsprechen solche Testsituationen insbesondere Klassenarbeiten und Klausuren. Die Interpretation dieser Testerfahrungen nimmt im Verlauf der Schulzeit eine wichtige Rolle für die Ausprägung der individuellen Kompetenzüberzeugungen ein (Brookhart, Walsh & Zientarski, 2006).

Das Erwartung-Wert-Modell der Motivation zur Testbearbeitung ist als zirkuläres Modell zu verstehen. Sowohl über verschiedene Testsituationen hinweg als auch innerhalb einer aktuellen Testsituation wird die Testleistung von der Testperson fortwährend wahrgenommen und interpretiert. Die motivationalen Überzeugungen und Dispositionen (z. B. Fähigkeitsselbstkonzept, Leistungsmotiv) sind dabei als relativ träge aufzufassen (vgl. Möller, 2008). Sie passen sich nicht direkt im Verlauf einer Testsituation an, sondern verändern sich im Verlauf der Zeit aufgrund von Erfahrungen im tagtäglichen Leistungserleben und über verschiedene Testsituationen hinweg. Dennoch beeinflussen sie natürlich die Motivation in einer aktuellen Testsituation. Die Erwartungs- und die Wertkomponente, die Testleistung und die Wahrnehmung und Interpretation der Testleistung werden hingegen während einer Testsituation fortwährend aktualisiert und adaptiert. Bei Anwendung des Modells auf die Testbearbeitung lässt sich der Motivationsprozess theoretisch demnach in folgende Schritte unterteilen:

1. Motivationale Überzeugungen und Dispositionen, die durch früher gemachte Erfahrungen in ähnlichen Testsituationen geprägt sind, werden in Form von Kompetenzüberzeugungen und affektgeprägten Erinnerungen a priori in die Testsituation mitgebracht.

2. Die sich daraus ergebende aktuelle Motivation zur Testbearbeitung wirkt sich, insbesondere in Form der subjektiven Erfolgserwartung, auf die aktuelle Testleistung aus.

3. Diese aktuelle Testleistung wird durch die Testperson umgehend wahrgenommen und interpretiert.

4. Daraufhin kann sich die aktuelle Motivation zur Testbearbeitung – abhängig davon, ob sich eine positive oder eine negative Diskrepanz zwischen der wahrgenommenen Leistung und den individuellen Kompetenzüberzeugungen ergibt – positiv oder negativ verändern, also zu- oder abnehmen.

5. Diese gegebenenfalls veränderte aktuelle Motivation zur Testbearbeitung wirkt sich erneut auf die aktuelle Testleistung aus, welche wiederum subjektiv wahrgenommen und interpretiert wird (Wiederholung der Schritte 2 bis 5).

Dieser Prozess setzt sich so lange fort, bis die Testsituation beendet ist. Die in verschiedenen Testsituationen gemachten Erfahrungen beeinflussen mittel-/langfristig die motivationalen Überzeugungen und Dispositionen (vgl. Schritt 1).

Die Zusammenschau macht deutlich, dass das Erwartung-Wert-Modell der Motivation zur Testbearbeitung State- und Trait-Aspekte der Leistungsmotivation berücksichtigt (vgl. Abschnitt 3.1). Zudem erlaubt die reziproke, komplexe und dynamische Konstruktion des Modells nicht nur eine Anwendung auf Makroebene (z. B. Vorhersagen der Motivation zur Testbearbeitung über Situationen hinweg), sondern auch auf Mikroebene (z. B. Vorhersagen der Motivation zur Testbearbeitung innerhalb einer Testsituation; vgl. auch Wolf, Smith & Birnbaum, 1995).

# 4 Empirischer Forschungsstand zur Motivation zur Testbearbeitung

In den Abschnitten 2 und 3 wurden die theoretischen Grundlagen zu computerisiertem adaptivem Testen und zu Motivation zur Testbearbeitung als einer spezifischen Form der Leistungsmotivation erläutert. Diese Informationen ermöglichen eine theoretisch fundierte, kritische Auseinandersetzung mit bisherigen empirischen Befunden zur Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests, die in diesem Abschnitt vorgenommen wird. Die Diskussion der empirischen Befunde mündet schließlich in die Ableitung der Fragestellungen und Hypothesen der vorliegenden Arbeit (Abschnitt 5).

Das Verständnis von Motivation zur Testbearbeitung als Bereitschaft, sich im Bearbeiten von Testaufgaben zu engagieren und hierfür Ausdauer und Anstrengung zu investieren (Baumert & Demmrich, 2001; vgl. Abschnitt 3.1), weist auf das Problem hin, dass die gezeigte Leistung in einem Test nicht notwendigerweise der maximalen Leistung der Testperson entspricht. Dennoch wird genau dies in vielen Anwendungen von Leistungstests implizit angenommen und das Testergebnis entsprechend als maximale Leistung interpretiert und bewertet (Wise, 2009; Wolf & Smith, 1995). Entspricht die gezeigte Leistung beispielsweise aufgrund von mangelnder Motivation zur Testbearbeitung nicht der maximalen Leistung, bedeutet dies folglich eine Beeinträchtigung der Validität, also der Gültigkeit der Interpretation der Testergebnisse als maximale Leistung (z. B. Dweck, Mangels & Good, 2004; Eklöf, 2008; Messick, 1995; Oakland & Harris, 2009; Thelk, Sundre, Horst & Finney, 2009). Daher ist es wichtig, bei der Interpretation von Testergebnissen die psychologischen Effekte eines Tests zu berücksichtigen (Sundre & Kitsantas, 2004). Die folgenden beiden Abschnitte stellen überblicksartig Ergebnisse zum Zusammenhang zwischen der Motivation zur Testbearbeitung und der Testleistung sowie zu Einflussfaktoren auf die Motivation zur Testbearbeitung in nicht-adaptiven Testverfahren (Abschnitt 4.1) und adaptiven Testverfahren (Abschnitt 4.2) dar. Anschließend werden die Befunde zusammenfassend diskutiert (Abschnitt 4.3).

## 4.1 Motivation zur Testbearbeitung in nicht-adaptiven Tests

In diesem Abschnitt wird die Motivation zur Testbearbeitung in herkömmlichen, nicht-adaptiven Testverfahren betrachtet. Mit herkömmlich beziehungsweise nicht-adaptiv ist gemeint, dass die zu bearbeitenden Aufgaben vor Beginn des Tests feststehen und sich die Auswahl der Testaufgaben somit nicht an das individuelle Antwortverhalten anpasst. Die Struktur dieses Abschnitts orientiert sich an der Struktur des Grundmodells motivierten Verhaltens in einer Testsituation (Abbildung 3.1). Zunächst werden Befunde zum Einfluss der Motivation zur Testbearbeitung auf die Testleistung dargestellt (Abschnitt 4.1.1). Dann werden Ergebnisse zu dem Einfluss von Personenmerkmalen (Abschnitt 4.1.2), Situationsmerkmalen (Abschnitt 4.1.3) und Testmerkmalen (Abschnitt 4.1.4) auf die Motivation und die Leistung berichtet. Es wird auf divergierende empirische Befunde eingegangen (Abschnitt 4.1.5) und schließlich ein kurzes Fazit gezogen (Abschnitt 4.1.6).

### 4.1.1 Motivation zur Testbearbeitung und Testleistung

Unter theoretischer Bezugnahme auf das Erwartung-Wert-Modell der Leistungsmotivation (Abschnitt 3.2.2) bestätigen zahlreiche empirische Studien konsistent einen positiven Einfluss der Motivation zur Testbearbeitung auf die Testleistung in Low-Stakes-Tests (z. B. Arvey, Strickland, Drauden & Martin, 1990; Cole et al., 2008; Eklöf, 2008; Kim & McLean, 1995; Pintrich, Smith, Garcia & McKeachie, 1993; Sundre, 1999; Sundre & Kitsantas, 2004; Thelk et al., 2009; Wolf & Smith, 1995; Wolf et al., 1995; zur Unterscheidung zwischen so genannten High- und Low-Stakes-Tests siehe Abschnitt 4.1.3). In vielen der genannten Studien wird Motivation zur Testbearbeitung über die selbstberichtete Anstrengung operationalisiert, die teils als ein der Erwartungskomponente verwandtes Merkmal (Wolf et al., 1995) und teils als Moderatorvariable zwischen Wertkomponente und Testleistung (z. B. Sundre, 1999) angesehen wird. Einige Studien berücksichtigen zusätzlich Wichtigkeit und/oder Nützlichkeit als Teil der Wertkomponente (z. B. Eklöf, 2008; Sundre & Kitsantas, 2004). Pintrich et al. (1993) sowie Wolf et al. (1995) erfassen neben der Anstrengung die Erfolgserwartung als Teil der Motivation.

In ihrer Metaanalyse empirischer Studien zum Zusammenhang zwischen Motivation zur Testbearbeitung und Testleistung stellen Wise und DeMars (2005) fest, dass in 24 von 25 Fällen eine höhere Motivation mit besseren Leistungen verbunden ist. Die mittlere Effektstärke für die motivationsbezogenen Leistungsdifferenzen beträgt mit $d = 0.59$ mehr als eine halbe Standardabweichung. Auch bezogen auf nicht-experimentelle Untersuchungen berichten die Autoren von einem positiven Zusammenhang zwischen Motivation und Leistung. Sundre und Kitsantas (2004) bestätigen insbesondere bei anspruchsvollen Aufgaben wie Aufsätzen einen signifikanten Einfluss der Motivation zur Testbearbeitung auf die Testleistung (vgl. auch Sundre, 1999). Thelk et al. (2009) weisen auf die generelle Notwendigkeit hin, die Motivation zur Testbearbeitung zu prüfen, sofern mit dem Test die maximale Leistung gemessen werden soll. Mangelnde Motivation kann den Autorinnen zufolge als systematischer „Fehler“ interpretiert werden, der die Testleistung in der Regel negativ verzerrt.

Eklöf (2008) bestätigt in ihrer Analyse der Motivation zur Testbearbeitung in der groß angelegten Vergleichsstudie *Trends in International Mathematics and Science Study* (TIMSS) 2003 eine signifikante, wenn auch numerisch eher geringe Korrelation zwischen mathematiktestbezogener Motivation und Mathematiktestleistung ($r = .25$) in der schwedischen Teilstichprobe. Ein großer Anteil der Varianz der Mathematikleistung (42 %) lässt sich auf motivationale Variablen zurückführen. Das mathematikbezogene Fähigkeitsselbstkonzept, das hier als Teil der Erwartungskomponente operationalisiert wurde, nimmt dabei die bedeutsamste Rolle ein. Doch auch die testsituationsspezifische Motivation im Sinne der Wichtigkeit der bestmöglichen Bearbeitung der Testaufgaben und der dabei aufgewendeten Anstrengung leistet einen signifikanten inkrementellen Beitrag zur Varianzaufklärung der Mathematikleistung. Der allgemeine Wert, der Mathematik beigemessen wird, zeigt keinen eigenständigen Beitrag zur Varianzaufklärung, was mit den theoretischen Annahmen aus dem Erwartung-Wert-Modell von Eccles und Wigfield (2002) vereinbar ist. Absolut betrachtet bescheinigen sich die in TIMSS 2003 getesteten Achtklässlerinnen und Achtklässler in Schweden eine recht hohe Motivation zur Testbearbeitung, obwohl es sich um einen Test handelt, bei dem die Teilnahme nicht mit individuellen Konsequenzen verbunden ist. Interviews mit einigen der beteiligten Schülerinnen und Schülern lassen eine hohe extrinsische Motivation erkennen, Schweden in der Studie würdig zu vertreten und so ein gutes Ergebnis im Vergleich zu anderen Staaten zu erzielen. Baumert und Demmrich (2001) vermuten eine vergleichbar

motivierende Wirkung der Testinstruktion von PISA 2000 in der deutschen Teilstichprobe, da die Instruktion die Bedeutsamkeit der Testteilnahme für die Staaten aus aller Welt herausstellt. Die Autoren ziehen den Schluss, dass die Anlage solch großer internationaler Studien bereits ausreichend Interesse beziehungsweise intrinsische Motivation in den teilnehmenden Personen auslöst, um die Testaufgaben angestrengt und ausdauernd zu bearbeiten.

Schmidt-Atzert (2006) weist darauf hin, dass eine Untersuchung der Leistungsmotivation in Testsituationen den reziproken Zusammenhang zwischen Leistungsmotivation und Leistung angemessen berücksichtigen muss. Er spricht sich für die Verwendung von Wiederholungsmessungen in solchen Studien aus, um den prozesshaften Charakter und den zeitlichen Verlauf von Motivation und Leistung abbilden zu können (vgl. auch Hambleton et al., 1991; Schunk et al., 2008).

### 4.1.2 Personenmerkmale, Motivation zur Testbearbeitung und Testleistung

Schunk et al. (2008) und Wigfield (1994) weisen unter Bezugnahme auf Studien aus Eccles' Arbeitsgruppe darauf hin, dass Schülerinnen und Schüler mit einer positiven Selbstwahrnehmung ihrer Kompetenz und mit positiven Erfolgserwartungen auch bei Kontrolle der Ausgangsleistungen bessere Leistungen zeigen, sich mehr anstrengen und ausdauernder sind als andere Schülerinnen und Schüler. Zur Vorhersage der Leistung (in Form von Zeugnisnoten) erweisen sich die Kompetenzüberzeugungen und die Erfolgserwartung in diesen Studien sogar als bedeutsamer als die vorherige Leistung.

In Bezug auf das dispositionelle Leistungsmotiv zeigt sich empirisch, wie theoretisch zu erwarten (vgl. Abschnitt 3.2), lediglich ein indirekter Einfluss auf die tatsächliche Testleistung (Krau, 1982; zitiert nach Brunstein & Heckhausen, 2006, S. 171; Kuhl, 1983; Rink, 1994). Dementsprechend ist häufig auch nur ein schwacher Einfluss des Leistungsmotivs auf Schulnoten festzustellen (Weiner, 1994). Allerdings führt Weiner für diesen Befund noch als anderes Argument an, dass Noten einen überdeterminierten Leistungsindex darstellen, der von vielen verschiedenen Motiven zugleich beeinflusst wird. Schmalt und Langens (2009) heben hervor, dass für das Wirksamwerden des Leistungsmotivs entsprechende anregende situative Bedingungen (z. B. im Klassenzimmer) gegeben sein müssen, da das Motiv allein aus sich heraus keinen Einfluss auf die Schulleistung nimmt. Nicholls (1984, zitiert nach Brunstein & Heckhausen, 2006, S. 149) weist darauf hin, dass stark misserfolgsmotivierte Personen (d. h. Personen mit einer hohen Ausprägung der leistungsvermeidenden Disposition Furcht vor Misserfolg; vgl. Abschnitt 3.2) die subjektive Schwierigkeit von Aufgaben höher einschätzen als gering misserfolgsmotivierte Personen.

Kuhl (1983) referiert empirische Befunde, dass sich insbesondere im Falle einer großen Diskrepanz zwischen erwartetem Erfolg und wahrgenommenem Erfolg eine sukzessive, deutliche Anpassung der Erwartungskomponente (im Sinne des Fähigkeitsselbstkonzepts) zeigt.

### 4.1.3 Situationsmerkmale, Motivation zur Testbearbeitung und Testleistung

Im Hinblick auf den Zusammenhang zwischen Motivation zur Testbearbeitung und Testleistung scheint es wichtig, das Situationsmerkmal der Bedeutsamkeit der Testergebnisse zu berücksichtigen (vgl. Abschnitt 3.1; z. B. Wainer, 1993): Nach Wise (2009) sind so genannte *Low-Stakes*-Testsituationen dadurch gekennzeichnet, dass das Abschneiden im Test keinerlei persönliche Konsequenzen für die Testperson hat; sie hat also weder Vor- noch Nachteile in Abhängigkeit ihres Testergebnisses zu erwarten. Ein Beispiel für Low-Stakes-Tests im Bildungsbereich ist PISA (z. B. OECD, 1999). Im Gegensatz dazu ist das Abschneiden in *High-Stakes*-Testsituationen mit individuellen Konsequenzen für die Testperson verbunden. Diese Situation liegt beispielsweise vor, wenn das Testergebnis die Zulassung zu einem Studiengang bestimmt (wie z. B. beim GMAT in den USA; vgl. Abschnitt 2.4) oder wenn der Test, wie im Falle einer Schulklausur, benotet wird und die Note in ein Zeugnis eingeht.

Empirisch zeigt sich grundsätzlich das Bild, dass in einer Low-Stakes-Testsituation eine größere Motivationsvarianz zwischen den Personen zu beobachten ist, während in einer High-Stakes-Testsituation in der Regel nahezu alle Testpersonen hoch motiviert sind (vgl. Smith & Smith, 2002; Sundre & Kitsantas, 2004). Daher scheint die Motivation zur Testbearbeitung in High-Stakes-Tests für Unterschiede in der Testleistung kaum eine Rolle zu spielen; stattdessen nimmt hier Testängstlichkeit eine größere Bedeutsamkeit ein (Sundre & Kitsantas, 2004). Wolf und Smith (1995) berichten eine signifikant niedrigere Motivation zur Testbearbeitung und Testleistung unter Low-Stakes-Bedingungen im Vergleich zu High-Stakes-Bedingungen (mit einer sehr großen Effektstärke von $d = 1.45$ bzgl. Motivation und einer kleinen Effektstärke von $d = 0.26$ bzgl. Leistung; vgl. auch Arvey et al., 1990; Sundre & Kitsantas, 2004). Der Zusammenhang zwischen Motivation und Leistung ist mit $r = .35$ und $r = .23$ in beiden Testbedingungen signifikant und von mittlerer Höhe. Die Autoren empfehlen vor diesem Hintergrund, bei der Verwendung von Testnormen zur Interpretation von Testergebnissen zu berücksichtigen, ob die Normen auf der Basis von Low- oder High-Stakes-Testbedingungen erstellt wurden. Unter Low-Stakes-Bedingungen generierte Testnormen für die Bewertung von Ergebnissen aus regelmäßigen Leistungskontrollen im staatlichen Schulsystem (High-Stakes-Bedingungen) zu verwenden, wie in den USA durchaus üblich (vgl. auch Wise & DeMars, 2005), kann zu einer invaliden Interpretation der tatsächlichen Leistung der Schülerinnen und Schüler führen.

Die Differenzen in Abhängigkeit von der Bedeutsamkeit der Testergebnisse sind damit zu erklären, dass in einer High-Stakes-Testsituation insbesondere die Nützlichkeit und die Wichtigkeit (als Teilaspekte der Wertkomponente) sowie die Anstrengung einen positiven Zusammenhang mit der Testleistung zeigen (z. B. Thelk et al., 2009). Die Möglichkeit, durch ein gutes Testergebnis angestrebte Ziele zu erreichen, fördert die Anstrengung im Test und die subjektive Wichtigkeit, so dass die gezeigte Leistung in der Regel bei allen Testpersonen eher der maximalen Leistung entspricht (Sundre & Kitsantas, 2004). Verzerrungen der Testergebnisse aufgrund einer unterschiedlich hohen Motivation zur Testbearbeitung sind unter High-Stakes-Testbedingungen weniger wahrscheinlich. Sind hingegen keinerlei persönliche Konsequenzen der Testergebnisse zu erwarten, geht der Einfluss der Nützlichkeit zurück und die Erwartungskomponente der Leistungsmotivation gewinnt für die Prädiktion der Testleistung an Bedeutung. Da die Erfolgserwartung maßgeblich von den subjektiven Kompetenzüberzeugungen abhängt, sind hier spätestens ab der Adoleszenz stärkere interindividuelle Unterschiede und damit eine höhere

potentielle Verzerrung der resultierenden Testleistung zu erwarten. Insbesondere im Falle von Low-Stakes-Testsituationen ist es für eine valide Interpretation der Testergebnisse demnach wichtig, die Motivation zur Testbearbeitung zu beachten. Daher beziehen sich die Ausführungen in dieser Arbeit vorwiegend auf derartige Testsituationen.

### 4.1.4 Testmerkmale, Motivation zur Testbearbeitung und Testleistung

Thelk et al. (2009) führen an, dass eine motivierende, nicht bedrohlich formulierte Testinstruktion zu einer höheren Erfolgserwartung führt und dass sowohl die Erwartungs- als auch die Wertkomponente einen direkten Einfluss auf das Testverhalten in einer Testsituation nehmen. Bereits McClelland et al. (1953) vermuten, dass leistungsorientiert formulierte Testinstruktionen die validesten Testergebnisse im Sinne einer maximalen Testleistung ermöglichen, auch wenn diese die Gefahr beinhalten, Testängstlichkeit hervorzurufen. Lang und Fries (2006) bestätigen, dass sich aktivierende, leistungsorientiert formulierte Testinstruktionen dazu eignen, selbstattribuierte Motive wie Hoffnung auf Erfolg zu aktivieren, die wiederum situationsspezifisches, spontanes leistungsorientiertes Verhalten hervorrufen (vgl. auch Spangler, 1992). Auch Brown und Walberg (1993) berichten im Hinblick auf die Testleistung von Grundschülerinnen und -schülern einen signifikant positiven Effekt einer Testinstruktion, welche die Aufforderung beinhaltet, im Test das Beste zu geben und sich anzustrengen, für sich selbst, die Eltern und den Testleiter, gegenüber einer Testinstruktion, die diese Aufforderung nicht beinhaltet (durchschnittliche Effektstärke: $d = 0.30$).

Wise (2009) sowie Wise und DeMars (2005) machen verschiedene Vorschläge, wie die situative Motivation zur Testbearbeitung in Low-Stakes-Tests erhöht werden könnte. Die Vorschläge beziehen sich generell entweder darauf, mögliche Verzerrungen durch motivationale Effekte nachträglich im Rahmen der Datenanalyse zu berücksichtigen, oder darauf, am Leistungstest selbst anzusetzen und ihn so zu verändern, dass er möglichst motivierend wirkt. Ein Beispiel für die nachträgliche Korrektur der Daten um Motivationseffekte ist der Einbau eines so genannten Motivationsfilters: Die Daten der Personen, die eine geringe Motivation zur Testbearbeitung berichten, werden – unter der Voraussetzung, dass kein Zusammenhang zwischen Motivation und Leistungsausprägung besteht – schlichtweg eliminiert. Dies kann jedoch zu einem großen Datenverlust und zu einer systematischen Verzerrung der Ergebnisse führen und erscheint daher keine empfehlenswerte Strategie zu sein. In Bezug auf den anderen Ansatz, die Anpassung des Tests, nennen die Autoren unter anderem die Möglichkeit computerisierten adaptiven Testens: Diese Art zu testen sollte den Autoren zufolge aufgrund der Darbietung von für jede Testperson mittelschwierigen Aufgaben zu einer Maximierung der intrinsischen Motivation und damit auch zu Motivationsvorteilen gegenüber herkömmlichem Testen führen. Inwiefern bisherige empirische Befunde diese Vermutung unterstützen, wird in Abschnitt 4.2 erläutert.

### 4.1.5 Kontroverse Befunde zur Motivation zur Testbearbeitung in nicht-adaptiven Tests

Im Widerspruch zu vielen der dargestellten Befunde stehen Ergebnisse der Studie von Baumert und Demmrich (2001). Die Autoren untersuchen den motivationalen und leistungsbezogenen Effekt einer Steigerung der Bedeutsamkeit für Testpersonen in groß angelegten Vergleichsstudien durch Maßnahmen wie die Gewährung individuellen Feedbacks, die Benotung der Testergebnisse oder eine finanzielle Belohnung nach vollständiger Bearbeitung aller Testaufgaben und Fragebögen. Das Hauptergebnis ist, dass sich weder Motivation noch Testleistung durch die drei vorgenommenen Maßnahmen signifikant erhöhen ließen.

Wolf und Smith (1995) berichten anhand einer experimentellen Studie mit Psychologie-Studierenden gar einen stärkeren positiven Einfluss der Motivation zur Testbearbeitung auf die Testleistung (Multiple-Choice-Test zu Inhalten des Psychologie-Studiums) unter High- als unter Low-Stakes-Bedingungen (vgl. auch Smith & Smith, 2002). Eine Erklärung für dieses Ergebnis äußern die Autoren nicht. Der Befund konnte jedoch nicht repliziert werden (z. B. Sundre, 1999).

In einer Studie von O'Neil, Sugrue und Baker (1996) lassen sich Schülerinnen und Schüler in der achten Klassenstufe noch durch eine finanzielle Belohnung zu größerer Anstrengung und besserer Testleistung in einem Mathematiktest motivieren, jene in der 12. Klassenstufe hingegen nicht mehr. Motivierend formulierte Testinstruktionen zeigen in keiner der Klassenstufen einen Effekt auf die Anstrengung oder die Testleistung. Die Autoren vermuten, dass die Wirkung der Testinstruktion auf die Anstrengung und die Testleistung durch eine persönliche Beziehung zwischen Testpersonen und Testleitung bedingt sein könnte: Während die Studie von O'Neil et al. (1996) von neutralen Testleiterinnen und Testleitern durchgeführt wurde, geschah die Testadministration bei Brown und Walberg (1993; vgl. Abschnitt 4.1.4) durch die Lehrkräfte der getesteten Schulkinder.

Gagné und StPère (2001) finden keinerlei signifikante Zusammenhänge zwischen Leistungsmotivation und Leistung im Schulbereich. Allerdings sind die Generalisierbarkeit und die Bedeutsamkeit dieser Ergebnisse für die vorliegende Arbeit eingeschränkt: So untersuchen die Autoren eine reine Mädchenstichprobe einer kanadischen Privatschule (8. Klassenstufe). Zudem erfassen sie eine konzeptuell andere Leistungsmotivation (vgl. Deci & Ryan, 1993) und deren Auswirkungen über einen Zeitraum von mehreren Monaten.

### 4.1.6 Fazit zu den Befunden zur Motivation zur Testbearbeitung in nicht-adaptiven Tests

Der Großteil der empirischen Befunde zur Motivation zur Testbearbeitung bei nicht-adaptiven Tests unterstützt die Bedeutsamkeit von Personen- (z. B. Leistungsmotiv), Situations- (z. B. Bedeutsamkeit der Testergebnisse) und Testmerkmalen (z. B. Testinstruktion) für die resultierende Leistungsmotivation und die Testleistung (vgl. Grundmodell motivierten Verhaltens in einer Testsituation, Abschnitt 3.1). Zudem bestätigen die meisten Ergebnisse die theoretischen Annahmen des Erwartung-Wert-Modells der Motivation zur Testbearbeitung (Abschnitt 3.2.2.4). Dies gilt insbesondere für den positiven Zusammenhang zwischen aktueller Motivation zur Testbearbeitung und Testleistung, für die Rückwirkung der wahrgenommenen Leistung auf motivationale Faktoren

und für die enge Verbindung zwischen Erfolgserwartung und Testleistung sowie zwischen Wert und Wahl-/Entscheidungsverhalten. Es scheint systematische Einflussfaktoren auf die Motivation zur Testbearbeitung zu geben, die das Zeigen der maximalen Leistung im Test beeinträchtigen können.

## 4.2 Motivation zur Testbearbeitung in adaptiven Tests

Während es zur Motivation zur Testbearbeitung in nicht-adaptiven Tests zahlreiche empirische Studien gibt, liegen nur wenige Befunde zur Motivation zur Testbearbeitung bei adaptiven Tests vor. Dabei ist die Motivation zur Testbearbeitung auch im CAT prädiktiv bedeutsam für die Testleistung (Kim & McLean, 1995). Im Grundmodell motivierten Verhaltens in einer Testsituation (Abschnitt 3.1) wird Testmerkmalen wiederum ein Einfluss auf die Motivation zur Testbearbeitung zugesprochen. Dieser Einfluss zeigt sich auch empirisch (vgl. Abschnitt 4.1). Daher ist davon auszugehen, dass auch das Testmerkmal Testalgorithmus, nicht-adaptiv (FIT) versus adaptiv (CAT), Auswirkungen auf die Motivation zur Testbearbeitung und die Testleistung haben kann. In Abhängigkeit von der gewählten theoretischen Begründung lassen sich dabei sowohl motivierende als auch demotivierende Effekte von CAT ableiten. Nachfolgend werden zu beiden Annahmen theoretische Überlegungen und empirische Befunde präsentiert (Abschnitte 4.2.1 und 4.2.2) sowie einige weitere empirische Befunde zum Einfluss von Personen-, Situations- und Testmerkmalen auf die Motivation zur Testbearbeitung im CAT berichtet (Abschnitt 4.2.3).

### 4.2.1 Überlegungen und Befunde zu einer motivationssteigernden Wirkung von CAT

Theoretisch ist eine motivierende Wirkung von CAT im Vergleich zu FIT zum Beispiel aus dem der Lernforschung entstammenden Konzept der *Motivationalen Zone der proximalen Entwicklung* (Vygotsky, 1978; vgl. Brophy, 1999) ableitbar: Optimale, motivierende Aufgaben sollten so konzipiert sein, dass sie für die Testperson weder zu schwierig noch zu einfach sind. In einem FIT ist folglich die Wahrscheinlichkeit recht hoch, dass ein Teil der Testpersonen entweder entmutigt oder gelangweilt ist, weil die Testaufgaben viel zu schwierig oder viel zu einfach sind (Betz, 1975; Georgiadou et al., 2006). In einem CAT hingegen wäre auf Basis dieser Theorie von einer individuell optimalen Aufgabenauswahl auszugehen. Auch Hambleton et al. (1991) heben die Minimierung der Testfrustration einiger Testpersonen explizit als einen Vorteil von CAT hervor.

Empirisch liegen noch keine eindeutigen Befunde zur Motivation zur Testbearbeitung im CAT vor, auch wenn es Ergebnisse gibt, die die Annahme einer motivationssteigernden Wirkung von CAT stützen. So führten einige Studien in den 1970er Jahren zu den Auswirkungen von Testalgorithmus, Leistungsrückmeldung und allgemeiner Leistungsausprägung auf die Testleistung und die Motivation zur Testbearbeitung zu der jahrzehntelang vorherrschenden Auffassung, dass CAT motivationsförderlich wirkt (Betz, 1975; Betz & Weiss, 1976a, 1976b). Die Autoren berichten einen positiven Effekt von Leistungsrückmeldung (Information über die korrekte Antwort) auf die Leistung in einem konventionellen, nicht-adaptiven Vokabeltest. Dieser Effekt zeigt sich unabhängig vom Leistungsniveau der untersuchten Studierenden. In der adaptiven Version des Vokabeltests wirkt sich

die Leistungsrückmeldung nur bei hoch leistungsfähigen Testpersonen positiv auf die Leistung aus, bei den anderen Testpersonen bewirkt sie keinen Leistungsunterschied. Die weniger leistungsfähigen Testpersonen zeigen darüber hinaus im CAT generell eine deutlich bessere Leistung als im FIT (Betz, 1975; Betz & Weiss, 1976a). Betz und Weiss ziehen aus den Ergebnissen den Schluss, dass die Bearbeitung eines CAT für weniger fähige Personen ähnlich motivierend wirkt wie die Bereitstellung einer Leistungsrückmeldung im FIT. Diese wurde von 90 Prozent aller Testpersonen positiv aufgenommen, vorwiegend mit der Begründung, dass die Kenntnis der richtigen Antwort den Test interessanter macht. Die Autoren schreiben die motivierende Wirkung des CAT in der Teilstichprobe mit niedriger Leistungsausprägung der Tatsache zu, dass die eigene Leistung als unerwartet hoch wahrgenommen wird: Im CAT können diese Personen wie alle Personen etwa 50 Prozent der Aufgaben korrekt lösen (Betz & Weiss, 1976a: 47 %), während die entsprechende Prozentzahl in einem FIT in der Regel niedriger ausfällt (Betz & Weiss, 1976a: 40 %). Den ausbleibenden motivierenden Effekt des CAT im Vergleich zum FIT bei hoch leistungsfähigen Testpersonen erklären die Autoren entsprechend. So liegt ihr Anteil korrekt beantworteter Aufgaben im CAT bei 50 Prozent, im FIT hingegen bei 58 Prozent. Weshalb diese, numerisch sogar größere, Differenz zugunsten des FIT bei den hoch leistungsfähigen Personen im CAT jedoch nicht motivationsmindernd wirkt, erläutern Betz und Weiss nicht. Sie interpretieren ihre Befunde als Hinweis darauf, dass konventionelle Tests nicht notwendigerweise die maximale Leistung aller Testpersonen erfassen, was die Validität solcher Ergebnisse beeinträchtigen mag.

In einer vertiefenden Studie mit einem Teil derselben studentischen Stichprobe untersuchen Betz und Weiss (1976b) den Effekt von Testalgorithmus (CAT/FIT), Leistungsrückmeldung (ja/nein) und allgemeiner Leistungsausprägung (hoch/niedrig, basierend auf der College-Zugehörigkeit der Testpersonen) auf die Motivation zur Testbearbeitung und die Testängstlichkeit. Die Autoren operationalisieren Motivation über drei Fragen zur Anstrengung und Herausforderung im Test und eine Frage zur Wichtigkeit der Bearbeitung, leider ohne Angaben zur theoretischen Verankerung des Merkmals. Die Testängstlichkeit wird über drei selbst erstellte, an bestehende Fragebögen angepasste Fragen erfasst. Zudem liegen Einschätzungen der Testpersonen zur wahrgenommenen Testschwierigkeit vor. Nach der Bearbeitung des Vokabeltests erfolgen die Angaben zur Motivation und zur Testängstlichkeit und schließlich erneut die Bearbeitung eines Vokabeltests. Zentrale Ergebnisse der Studie sind:

- Hoch leistungsfähige Testpersonen zeigen generell eine höhere Motivation zur Testbearbeitung als weniger leistungsfähige Testpersonen.
- Die Haupteffekte von Testalgorithmus und Leistungsrückmeldung auf die Motivation sind hingegen nicht signifikant.
- Hoch leistungsfähige Testpersonen sind im CAT und im FIT ähnlich hoch motiviert, wohingegen weniger leistungsfähige Testpersonen im CAT signifikant motivierter sind als im FIT. Das heißt, nur bei den weniger leistungsfähigen Personen bewirkt der Testalgorithmus einen direkten Einfluss auf die Motivation.
- Die Leistungsrückmeldung wirkt sich bei hoch leistungsfähigen Personen positiv auf die Motivation aus, bei weniger leistungsfähigen Personen negativ.

- Die deutlich niedrigsten Motivationswerte zeigen sich bei weniger leistungsfähigen Personen im FIT, gleichermaßen mit und ohne Leistungsrückmeldung.
- Der Zusammenhang zwischen Motivation und Leistung ist mit $r$ = .21 eher klein, aber signifikant.
- CAT und FIT werden subjektiv nicht als signifikant unterschiedlich schwierig wahrgenommen, wobei die subjektive Schwierigkeitswahrnehmung im FIT eher mit der objektiven Testschwierigkeit übereinstimmt als im CAT.

Betz und Weiss (1976b) fassen zusammen, dass CAT bei weniger leistungsfähigen Personen zu einer Motivations- und Leistungssteigerung führt, während hoch leistungsfähige Personen unabhängig vom Testalgorithmus eine hohe Motivation und Leistung zeigen. Im CAT gelingt es jedoch signifikant schlechter als im FIT, die Schwierigkeit der Testaufgaben korrekt wahrzunehmen und das eigene Abschneiden im Test richtig einzuschätzen. Ihre Ergebnisse interpretieren Betz und Weiss so, dass CAT generell motivierend wirkt, einheitlichere psychologische Ausgangsbedingungen schafft als FIT und damit allen Testpersonen ermöglicht, ihre maximale Leistung zu zeigen. Die Autoren stellen jedoch die eingeschränkte Verallgemeinerbarkeit ihrer Ergebnisse heraus, da die Stichprobe ausschließlich aus Studierenden besteht, die verwendeten Fragebögen nicht etabliert sind und möglicherweise Verzerrungen aufgrund von sozialer Erwünschtheit vorliegen.

Trotz dieser selbst formulierten Einschränkungen und wegen fehlender anderer Studien prägten die Studien von Betz und Weiss über lange Zeit das Bild von CAT als besonders motivierende Art zu testen (z. B. Kubinger, 1995; Volz-Sidiropoulou, 2004). Dass das Risikowahl-Modell (Atkinson, 1957, 1964; vgl. Abschnitt 3.2.1) zumindest für erfolgsmotivierte Personen ein Motivationsmaximum bei Aufgaben mittlerer Schwierigkeit vorhersagt und im CAT allen Testpersonen Aufgaben mittlerer Schwierigkeit vorgegeben werden, mag zu einer Verfestigung dieses Eindrucks beigetragen haben. Allerdings sei kritisch angemerkt, dass sich das Risikowahl-Modell streng genommen gar nicht auf Leistungstests ohne Aufgabenwahlmöglichkeit anwenden lässt (vgl. auch Abschnitt 4.3). Auch in einem Vergleich der papierbasierten, nicht-adaptiven Version der *Armed Services Vocational Aptitude Battery* (ASVAB; vgl. Abschnitt 2.4) mit der computerisierten, adaptiven Version dieses Tests beschreiben sich die Testpersonen (Militärrekruten) im CAT als höher motiviert als im FIT, obwohl sie den FIT als einfacher wahrnehmen (Arvey et al., 1990).

### 4.2.2 Überlegungen und Befunde zu einer motivationsmindernden Wirkung von CAT

Für die Annahme einer niedrigeren Motivation im CAT als im FIT spricht theoretisch der Befund, dass es für verschiedene Situationen bestimmte Adaptationsniveaus gibt, die als „normal“ oder neutral empfunden werden (McClelland et al., 1953). Während kleinere Abweichungen von diesen Niveaus als positiv oder anregend wahrgenommen werden, erregen größere Abweichungen negative Reaktionen und Unlust. Bezogen auf eine Leistungstestsituation ist davon auszugehen, dass Testpersonen vorwiegend Erfahrung mit nicht-adaptivem Testen haben (vgl. Lunz & Bergstrom, 1994). FIT stellt somit die gewohnte situative Gegebenheit dar. Bearbeiten die Testpersonen erstmals einen CAT, entspricht diese Situation insbesondere für die Testpersonen mit extrem hoher oder

extrem niedriger Leistungsausprägung einer größeren Abweichung vom neutralen Adaptationsniveau, was zu Motivationsverlust führen kann. Zu einem ähnlichen Schluss kommt man bei Anwendung der *Theorie der kognitiven Dissonanz* von Festinger (1957, zitiert nach Pintrich & Schunk, 2002, S. 38; vgl. auch Breckler, Olson & Wiggins, 2006): Menschen streben danach, konsistente Zusammenhänge zwischen ihren Kognitionen, Meinungen und ihrem Verhalten zu bewahren. Treten Inkonsistenzen, zum Beispiel zwischen Kognitionen, auf, erlebt die Person Dissonanz. Es entsteht der Drang, diese Unstimmigkeiten wieder zu beheben, und zwar umso dringlicher, je wichtiger die Kognitionen empfunden werden. Behoben werden können die Inkonsistenzen beispielsweise durch Änderung der Kognitionen oder durch Abwertung der Wichtigkeit. Bezogen auf Motivation im CAT oder im FIT ist davon auszugehen, dass aufgrund der Neuartigkeit der Testsituation „CAT“ hier eher Diskrepanzen, beispielsweise zwischen erfolgsbezogenen Kognitionen (Erfolgserwartung) und wahrgenommenem Verhalten (z. B. wahrgenommene Testleistung), auftreten als im FIT. Denn üblicherweise liegt in Bezug auf die Details des Testalgorithmus Intransparenz vor, das heißt, die Testpersonen werden vor dem Test nicht über die Besonderheiten des Algorithmus aufgeklärt, der dem Test zugrunde liegt. Daher ist davon auszugehen, dass die Testpersonen einen CAT mit der Erwartungshaltung bearbeiten, die sie aufgrund ihrer Erfahrungen mit FIT ausgebildet haben. Dies führt jedoch bei einem Großteil der Testpersonen dazu, dass positive oder negative Inkonsistenzen zwischen der Erfolgserwartung und der wahrgenommenen Testleistung auftreten; nur ein kleiner Teil der Testpersonen hat vermutlich vor Testbeginn die Erwartung, die Hälfte der Testaufgaben lösen zu können. Dieser Dissonanz-Zustand sollte als aversiv erlebt werden und könnte dazu führen, dass entweder die Erfolgserwartung oder die Wichtigkeit der Testergebnisse (Wertkomponente) herabgesetzt wird und somit eine Motivationsreduktion im CAT stattfindet (vgl. auch Abschnitt 3.2.2.1 zur möglichen Beeinträchtigung der Motivation durch eine Diskrepanz zwischen objektiver und subjektiver Erfolgserwartung). Dabei ist die Dissonanz umso größer und der Drang nach Auflösung der Dissonanz umso stärker, je gravierender die Inkonsistenz ist und je bedeutsamer die betroffenen Kognitionen für die Testperson sind.

Ferner stellt ein adaptiver Test eine sehr „schwache“ Situation dar in dem Sinne, dass das mögliche Verhalten minimal strukturiert beziehungsweise vorgegeben wird: Jede Testperson hat im CAT, nachdem einige Aufgaben bearbeitet wurden, für jede weitere Aufgabe eine mittlere objektive Erfolgswahrscheinlichkeit von 50 Prozent. Das heißt, es besteht maximale Unsicherheit in Bezug auf den Ausgang des gezeigten Verhaltens (Schmitt, Baumert & Hofmann, 2007). Gerade in schwachen Situationen werden Personenmerkmale wie das Fähigkeitsselbstkonzept oder die Selbstwirksamkeitserwartung stark verhaltenswirksam, was die Motivation zur Testbearbeitung in einem CAT beeinträchtigen könnte. Linacre (2000) vermutet, dass hoch leistungsfähige Testpersonen einen CAT als „traumatisches Erlebnis“ (S. 27) wahrnehmen, da ihre wahrgenommene Leistung weit hinter ihrer Erfolgserwartung zurückbleibt. Häufig ist im CAT bereits die erste Aufgabe schwieriger als im FIT, da für diese oft eine mittlere Schwierigkeit gewählt wird (Tonidandel, Quinones & Adams, 2002; vgl. Abschnitt 2.2), während die Eingangsaufgabe in einem FIT üblicherweise eine geringe Schwierigkeit aufweist („Eisbrecher-Item“; vgl. Lunz & Bergstrom, 1994). Zusammen mit anderen möglichen ungewohnten Testmerkmalen wie der fehlenden Möglichkeit des Zurückblätterns im CAT mag dies zu einer Verringerung der Motivation zur Testbearbeitung führen (Lunz et al., 1994). Aus diesen Gründen ist es unabdingbar, neben Simulationsstudien zu CAT auch Studien mit realen Testpersonen durchzuführen, so dass Effekte von CAT auf die Motivation zur Testbearbeitung abgeschätzt werden können.

Ergebnisse aktuellerer empirischer Studien zur Motivation zur Testbearbeitung bei adaptiven Tests bestätigen eine motivationsmindernde Wirkung von CAT: Frey, Hartig und Moosbrugger (2009) geben einer vorwiegend studentischen Stichprobe entweder die adaptive oder die nicht-adaptive Version eines Konzentrationsleistungstests vor (FAKT; Moosbrugger & Heyden, 1997) und setzen nach der Bearbeitung einiger Beispielaufgaben, aber vor der eigentlichen Testbearbeitung, den Fragebogen zur aktuellen Motivation ein (FAM; Rheinberg, Vollmeyer & Burns, 2001). Im CAT zeigen die Testpersonen eine signifikant niedrigere Motivation als im FIT. Dieser Effekt lässt sich ausschließlich auf die motivationale Teilkomponente Erfolgswahrscheinlichkeit zurückführen, das heißt, die Testpersonen antizipieren nach der Bearbeitung der Beispielaufgaben eine niedrigere subjektive Erfolgswahrscheinlichkeit im CAT als im FIT.

Der Befund einer niedrigeren Erfolgserwartung im CAT als im FIT passt gut zu der aus dem Erwartung-Wert-Modell ableitbaren Annahme, dass in einer Testsituation ohne Wahlmöglichkeit insbesondere die Erwartungskomponente Einfluss auf die aktuelle Motivation zur Testbearbeitung nimmt (vgl. Abschnitt 3.2.2.4). Die niedrigere Erfolgserwartung im CAT lässt sich wie folgt erklären: Jede Testperson kann im CAT, unabhängig von ihrem Leistungsniveau, 50 Prozent der vorgelegten Aufgaben korrekt beantworten (vgl. Abschnitt 2). Die Testpersonen wissen nicht, dass die Schwierigkeiten der vorgegebenen Aufgaben an das individuelle Leistungsniveau angepasst werden und dies sich auch im Testergebnis niederschlägt. Nach Tonidandel et al. (2002) orientiert sich die wahrgenommene Leistung im Test an der Anzahl richtig gelöster Aufgaben, ohne dass deren Qualität in Betracht gezogen wird. Folglich ist zu vermuten, dass die wahrgenommene Testleistung im CAT niedriger ausfällt als im FIT. Die wahrgenommene Testleistung beeinflusst wiederum die Erfolgserwartung (vgl. Abschnitt 3.2.2), daher ist wenig überraschend, dass die Erfolgswahrscheinlichkeit im CAT von der hoch leistungsfähigen Personengruppe niedriger eingeschätzt wird als im FIT. Denn hoch leistungsfähige Personen sind es in der Regel gewohnt, mehr als die Hälfte der Aufgaben korrekt lösen zu können. Der Zusammenhang zwischen der Anzahl korrekt gelöster Aufgaben und der Motivation zur Testbearbeitung wird vollständig durch die wahrgenommene Testleistung mediiert (Tonidandel et al., 2002). Diese Befunde regen an, durch eine Erhöhung der durchschnittlichen Lösungswahrscheinlichkeit im CAT von 50 Prozent auf etwa 70 Prozent mögliche Motivationseinbußen im CAT zu verhindern und so die wahrgenommene Zufriedenheit und Testfairness zu erhöhen (Andrich, 1995; Tonidandel et al., 2002). Eine solche Manipulation des adaptiven Testalgorithmus wäre auch aus psychometrischer Sicht noch akzeptabel, ohne allzu große Messeffizienzeinbußen in Kauf nehmen zu müssen (Bergstrom, Lunz & Gershon, 1992; Eggen & Verschoor, 2006). Leistungsrückmeldungen im CAT mögen hingegen aufgrund der hohen Divergenz zwischen wahrgenommener und tatsächlicher Testleistung bei Personen mit über- oder unterdurchschnittlicher Leistungsausprägung demotivierend wirken (vgl. Ausführungen zur Theorie der kognitiven Dissonanz in diesem Abschnitt). Generell muss erwähnt werden, dass der prozentuale Anteil der durch die Testschwierigkeit und die wahrgenommene Testleistung aufgeklärten Varianz an Motivation zur Testbearbeitung mit etwa fünf bis neun Prozent relativ gering ist (vgl. Metaanalyse von Spangler, 1992; Schmidt-Atzert, 2006).

### 4.2.3 Weitere Personen-, Situations- und Testmerkmale, Motivation zur Testbearbeitung und Testleistung

Um eine mögliche motivationsmindernde Wirkung von CAT zu vermeiden, schlagen Häusler und Sommer (2008) alternativ zu einer allgemeinen Erhöhung der mittleren Lösungswahrscheinlichkeit vor, im Verlauf eines typischen CAT einige einfache Aufgaben einzustreuen, die nicht in die Schätzung des Personenparameters eingehen. Auf diese Weise erfolgt keine Minderung der Messpräzision, und der Test wird lediglich um wenige Aufgaben verlängert. Hebt man die Lösungswahrscheinlichkeit eines CAT hingegen generell an, stellen die Autoren eine Unterschätzung des Personenparameters von hoch leistungsfähigen Testpersonen ($\theta \geq 2$) fest. Sie merken jedoch an, dass dies auch von der Qualität des verwendeten Aufgabenpools abhängt.

Im Hinblick auf Testängstlichkeit beziehungsweise Misserfolgsfurcht zeigen sich in den hier aufgeführten Studien keinerlei Unterschiede zwischen CAT und FIT. Allerdings steht eine hohe Testängstlichkeit im CAT in einem negativen Zusammenhang zur Testleistung. Im FIT zeigt sich kein Leistungsunterschied in Abhängigkeit von der Testängstlichkeit (Ortner & Caspers, in press). Eine Erläuterung der Aufgabenauswahl im CAT vor Beginn des Tests, also eine transparente Testinstruktion, führt zu einer höheren Testleistung (Ortner & Caspers, in press).

## 4.3 Diskussion der bisherigen Befunde zur Motivation zur Testbearbeitung

Die Zusammenschau der empirischen Befunde zur Motivation zur Testbearbeitung verdeutlicht, dass es angeraten ist, die Motivation der Testpersonen bei der Interpretation von Leistungstestergebnissen zu berücksichtigen. Denn Leistungstests zielen in der Regel auf die Erfassung der maximalen Leistung ab (Cronbach, 1970). Diese wird jedoch nur gezeigt, wenn die Testpersonen motiviert sind und sich bei der Bearbeitung der Testaufgaben anstrengen (vgl. Thelk et al., 2009). Es kommt nicht nur auf das Können, sondern auch auf das Wollen an (Pintrich & De Groot, 1990; Schmidt-Atzert, 2006). Eine mangelnde Motivation zur Testbearbeitung führt zu einer niedrigeren Testleistung und mindert folglich die Validität der Ergebnisse (Wise & DeMars, 2005). Die Gefahr der Verzerrung der Testergebnisse durch Motivationseinbußen ist bei Low-Stakes-Tests besonders hoch (Sundre & Kitsantas, 2004).

Baumert und Demmrich (2001) weisen darauf hin, dass eine ausreichende Motivation bei groß angelegten Vergleichsstudien wie PISA unerlässlich ist, allein schon um die Vergleichbarkeit der Ergebnisse zwischen den Staaten zu gewährleisten. Auch wenn die Teilnahme an internationalen Vergleichsstudien in der Regel unter Low-Stakes-Bedingungen für die Jugendlichen erfolgt, kann man auf aggregierter Ebene, wie auf Schul-, Bundesland- oder Staatenebene, von High-Stakes-Bedingungen sprechen (vgl. Cole et al., 2008; Sundre & Kitsantas, 2004). Gerade in Deutschland werden beispielsweise die Ergebnisse aus den PISA-Studien sehr intensiv diskutiert, und sie haben in den vergangenen Jahren mannigfaltige politische Reformen des Bildungssystems ausgelöst. In den USA ist seit Einführung des *No Child Left Behind Acts* (NCLB) im Jahr 2001 die valide Interpretation von Testergebnissen im Zusammenhang mit computerisiertem adaptivem Testen und dessen

regelmäßigem Einsatz zum Monitoring der Leistung staatlicher Schulen ein kontrovers diskutiertes Thema im Bereich der öffentlichen Bildung (z. B. Way, 2010; Wolf et al., 1995).

Trotz dieser offenkundigen Bedeutsamkeit von Motivation zur Testbearbeitung gibt es bislang, vor allem verglichen mit Untersuchungen zum überdauernden, generellen Leistungsmotiv, nur wenige Studien zur situationsspezifischen Leistungsmotivation (Eklöf, 2008; Rink, 1994). Eklöf hält die bestehenden Studien zudem für theoretisch und methodisch zu heterogen, als dass die Ergebnisse miteinander vergleichbar wären. Es muss festgestellt werden, dass die in einer Testsituation ablaufenden motivationalen Prozesse noch nicht hinreichend verstanden sind (vgl. Abschnitt 3.1). Die theoretischen Darlegungen zum Erwartung-Wert-Modell der Motivation zur Testbearbeitung (Abschnitt 3.2.2.4) legen nahe, die Motivation in einer Testsituation prozessorientiert zu betrachten. Motivation ist als adaptives System zu verstehen, das frühere Erfahrungen interpretiert und restrukturiert, um zukünftiges Handeln zu lenken (Middleton & Toluk, 1999). Motivation erfüllt damit zugleich einen antizipierenden und einen evaluierenden Zweck: antizipierend in der Form, dass Erfahrungen eine vorläufige Erwartungshaltung in Bezug auf die Erfolgswahrscheinlichkeit bei der Bearbeitung einer Aufgabe formen, die das Verhalten bestimmt, und evaluierend in der Form, dass die Erfahrungen, die eine Person in einer konkreten Situation macht, mit der „Schablone“ abgeglichen werden, welche die Person in ihrem Selbstsystem aufgrund ihrer früheren Erfahrungen angefertigt hat. In bekannten Situationen richten Personen ihr aktuelles Verhalten an früheren Erfahrungen aus, und zwar umso stärker, je vertrauter die Situation für sie ist. Motivation zur Testbearbeitung entsteht demnach aus der Interaktion zwischen Testperson, Testsituation und Testaufgabe, so dass die Berücksichtigung all dieser Merkmale wichtig ist (vgl. Abschnitt 3.1).

Zur Analyse der motivationalen Prozesse in einer Testsituation haben sich die theoretisch etablierten Erwartung-Wert-Modelle der Leistungsmotivation auch empirisch bewährt. Daher beruht der Großteil der referierten Studien auf Erwartung-Wert-Modellen. Zu kritisieren ist allerdings, dass der prozessuale Charakter häufig vernachlässigt wurde und keine Wiederholungsmessungen vorgenommen wurden. Außerdem beruhen nahezu alle Studien auf studentischen Stichproben. Dies erscheint vor dem Hintergrund problematisch, dass kognitive Selbstschemata wie das Fähigkeitsselbstkonzept sowie frühere Leistungserfahrungen einen zentralen Einfluss auf die Motivation zur Testbearbeitung ausüben. So ist davon auszugehen, dass Studierende über ein recht hohes allgemeines Leistungsniveau verfügen und folglich im Laufe ihres Lebens vorwiegend positive Leistungserfahrungen gesammelt haben. Wenn solche hoch selektiven Stichproben wie bei Betz und Weiss (1976a, 1976b) in „hoch leistungsfähige“ und „weniger leistungsfähige“ Personen eingeteilt werden, ist diese Etikettierung angesichts des insgesamt hohen allgemeinen Leistungsniveaus der untersuchten Stichprobe zu relativieren. Es ist nicht anzunehmen, dass die Ergebnisse aus solchen Studien ohne Weiteres auf breitere Grundgesamtheiten verallgemeinerbar sind. In der Studie von Betz und Weiss erscheint darüber hinaus wenig überzeugend, die College-Zugehörigkeit der Testpersonen als Kriterium für die Leistungsausprägung zu verwenden. Gerade die theoretische Einbettung in Erwartung-Wert-Modelle der Leistungsmotivation empfiehlt also die Verwendung breit angelegter Stichproben, auch, um generalisierbare Ergebnisse erhalten zu können. Dies gilt für die Analyse motivationaler Prozesse in konventionellen Tests, erst recht aber in adaptiven Tests. Zum einen entfaltet CAT seine psychometrischen Vorzüge vor allem in leistungsheterogenen Stichproben und kann gerade dann zu enormen Messeffizienz-Steigerungen führen, zum anderen werden differentielle Effekte des Testalgorithmus auf die Motivation zur Testbearbeitung erst in solchen

Stichproben sichtbar. Dennoch beruhen leider nahezu alle vorhandenen empirischen Studien zu Motivation im CAT und auch viele der Studien zu FIT auf studentischen oder gymnasialen Stichproben (Betz, 1975; Betz & Weiss, 1976a, 1976b; Frey, 2006; Frey et al., 2009; Ortner & Caspers, in press; Tonidandel et al., 2002).

Generell fällt ein Resümee zur Motivation zur Testbearbeitung im CAT gegenwärtig noch schwer, da sich die vorliegenden Ergebnisse teilweise widersprechen. Positiv hervorzuheben ist, dass in den vergangenen Jahren überhaupt Studien mit realen Testpersonen durchgeführt wurden, nachdem sich die Forschung zu CAT jahrzehntelang auf Simulationsstudien und die Analyse der psychometrischen Eigenschaften dieser Art zu testen beschränkt hat (Wolf & Smith, 1995). Bei näherer Betrachtung der älteren Studien von Betz und Weiss (1976a, 1976b) und der aktuelleren Studie von Frey et al. (2009) fällt auf, dass ein Vergleich und dementsprechend ein zusammenfassendes Fazit aus den Studien aus folgenden Gründen nur schwer möglich ist: Der theoretische Hintergrund bei Betz und Weiss wird nicht näher erläutert und unterbindet daher konzeptuelle Vergleiche. Die Operationalisierung der Motivation zur Testbearbeitung erfolgt bei Betz und Weiss anhand von vier selbst erstellten Fragen, die inhaltlich eher der Anstrengung, der Herausforderung und dem Wertaspekt Wichtigkeit zuzuordnen sind. Bei Frey et al. hingegen wird Motivation über den FAM (Rheinberg et al., 2001) operationalisiert, der vor allem die Erwartungskomponente erfasst. Sowohl aus theoretischer als auch aus empirischer Sicht sind diese Studien daher kaum vergleichbar. Leistung wurde in den bestehenden Studien über einen adaptiven Vokabeltest (Betz & Weiss, 1976a, 1976b) beziehungsweise über einen adaptiven Konzentrationsleistungstest (Frey et al., 2009) operationalisiert. Inwiefern die Zusammenhänge zwischen Testalgorithmus und Motivation zur Testbearbeitung von der Art des eingesetzten Leistungstests abhängen, ist unklar.

Die theoretische Einbettung bei Frey (2006) ins Risikowahl-Modell von Atkinson (1957, 1964) erscheint vor dem Hintergrund fehlender Wahlmöglichkeiten und angesichts der bekannten Kritikpunkte an dem Modell (vgl. Abschnitt 3.2.1) suboptimal. Die Verwendung des Risikowahl-Modells macht eine empirische Berücksichtigung sowohl der State- als auch der Trait-Komponenten der Leistungsmotivation erforderlich. Dies ist jedoch bei Frey (2006) und Frey et al. (2009) nicht erfolgt, wie die Autoren selbst kritisch bemerken. Es schließt sich das Problem an, dass sich die Hypothese einer höheren Motivation im CAT als im FIT (aufgrund der mittleren individuellen Lösungswahrscheinlichkeit von 50 Prozent im CAT) aus dem Risikowahl-Modell nur für erfolgsmotivierte Personen ableiten lässt. Ohne Erfassung des Leistungsmotivs lässt sich die Hypothese demnach gar nicht prüfen. Fragwürdig erscheint diese Hypothesenformulierung zudem insofern als sich die Annahme des Motivationsmaximums bei erfolgsmotivierten Personen im Risikowahl-Modell auf die subjektive Erfolgswahrscheinlichkeit bezieht, die gerade im CAT nur selten mit der objektiven Schwierigkeit übereinstimmt (vgl. Abschnitte 3.2.1 und 4.2).

Die heterogene Operationalisierung der Motivation zur Testbearbeitung in den bestehenden Studien erschwert generell die Vergleichbarkeit der Ergebnisse. Studien, in denen sowohl die Erwartungs- als auch die Wertkomponente der Motivation erfasst werden, liegen kaum vor. Gerade angesichts der komplexen motivationalen Prozesse erscheint eine differenzierte Erhebung der Leistungsmotivation mit soliden, theoretisch eingebetteten Instrumenten jedoch dringend geboten (Ortner & Caspers, in press).

Abschließend ist festzuhalten, dass die Motivation zur Testbearbeitung einen bedeutsamen Prädiktor für die maximale Leistung im Test darstellt. Die genauen motivationalen Prozesse in einer Testsituation werden jedoch noch nicht hinreichend verstanden. Hierzu bedarf es Studien, die Motivation inhaltlich differenziert und im Handlungsverlauf erfassen. Die Wirkung adaptiven Testens auf die Motivation zur Testbearbeitung, auch in Abhängigkeit von anderen Personen-, Situations- und Testmerkmalen, ist noch unklar. In Bezug auf die Untersuchung von Motivation zur Testbearbeitung im CAT erscheinen vor dem Hintergrund der theoretischen Überlegungen und angesichts der bestehenden empirischen Ergebnisse zwei Ansätze besonders viel versprechend: Zum einen sollte die Auswirkung einer Aufklärung über die Besonderheiten dieses Testalgorithmus vor Beginn des Tests auf die Motivation zur Testbearbeitung, also der Effekt einer transparenten Testinstruktion auf die Motivation, analysiert werden. Zum anderen erscheint es interessant, die Auswirkung einer Verringerung der Testschwierigkeit zu untersuchen.

Angesichts der viel versprechenden Möglichkeit, die Messeffizienz durch CAT gerade bei groß angelegten Studien massiv erhöhen zu können, und angesichts der Gefahr, eine Unfairness des Tests und Validitätseinschränkungen durch unangemessene Interpretationen der Testergebnisse zu riskieren, erscheint eine gründliche Analyse der Motivation zur Testbearbeitung in nicht-adaptiven und adaptiven Testverfahren angezeigt. Dies möchte die vorliegende Arbeit leisten. Im folgenden Abschnitt werden, unter Berücksichtigung der theoretischen und empirischen Grundlagen, die Fragestellungen und darauf aufbauende Hypothesen vorgestellt, mit denen diese Zielsetzung verfolgt wird.

# 5 Fragestellungen und Hypothesen

Die vorliegende Arbeit entsteht vor einem theoretischen und einem praktischen Beweggrund. Der *theoretische Beweggrund* ergibt sich daraus, dass es trotz zahlreicher theoretischer Ansätze bislang kein Modell zu den motivationalen Prozessen gibt, die in einer Testsituation ablaufen (Abschnitt 3.1). Die vorliegende Arbeit leistet hierzu einen ersten Schritt, indem sie den hierfür relevanten Teil des in Abschnitt 3.2.2.1 dargestellten Erwartung-Wert-Modells der Leistungsmotivation (Eccles & Wigfield, 2002) für Testsituationen spezifiziert (vgl. Abschnitt 3.2.2.4). Die erste Fragestellung lautet: *Eignet sich die Spezifikation des Erwartung-Wert-Modells der Leistungsmotivation von Eccles und Wigfield (2002), die Motivation zur Testbearbeitung in einer Leistungstestsituation zu erklären?* Das Ziel ist, Motivation zur Testbearbeitung angemessen in dem Modell abbilden zu können.

Der *praktische Beweggrund* beruht auf der Tatsache, dass computerisiertes adaptives Testen zwar aus psychometrischer Sicht eine sehr vielversprechende, messeffiziente Art des Testens ist, dass die psychologischen Effekte dieses Testalgorithmus auf die Motivation zur Testbearbeitung jedoch unklar sind. CAT bietet gerade für groß angelegte nationale oder internationale Vergleichsstudien eine Möglichkeit, langfristig Kosten zu senken und die Belastung für die Testpersonen zu verringern (Abschnitt 2). Zur psychologischen Wirkung von CAT gibt es jedoch widersprüchliche Befunde (Abschnitt 4.2). Eine umfassende, differenzierte und prozessorientierte Analyse der Motivation zur Testbearbeitung im CAT im Vergleich zum FIT steht noch aus. Die zweite Fragestellung dieser Arbeit ergibt sich aus diesen Überlegungen und lautet folgendermaßen: *Welches Bedingungsgefüge besteht zwischen Testmerkmalen (z. B. Testalgorithmus) und Personenmerkmalen (z. B. Fähigkeitsselbstkonzept) im Hinblick auf die Motivation zur Testbearbeitung?* Da der Anwendungskontext groß angelegter Vergleichsstudien wie PISA den Verzicht auf die Variation von Situationsmerkmalen legitimiert, beschränkt sich die vorliegende Studie auf Low-Stakes-Testsituationen. Das Ziel ist, eine Empfehlung geben zu können, wie ein computerisierter adaptiver Test konzipiert sein sollte, um zugleich eine hohe Messeffizienz und eine hohe Motivation zur Testbearbeitung zu gewährleisten.

Zu den beiden Fragestellungen werden im Folgenden Hypothesen abgeleitet. Die ersten vier Hypothesen beziehen sich auf die erste Fragestellung, die übrigen sechs Hypothesen betreffen die zweite Fragestellung. Zu jeder Hypothese wird zunächst kurz der Hintergrund erläutert und dann die Hypothese genannt.

Das Erwartung-Wert-Modell der Leistungsmotivation von Eccles und Wigfield (2002) wurde ursprünglich dafür entwickelt, Leistung und Wahlverhalten von adoleszenten Personen im Bereich mathematischer Kompetenz zu verstehen (Wigfield & Eccles, 2002). Eccles (2005) selbst erwähnt als Beispiel die Anwendung des Modells auf die Aufgabenbearbeitung. Daher erscheint das Modell geeignet, um es auf die Situation der Testbearbeitung zu spezifizieren und die motivationalen Prozesse im Verlauf der Testbearbeitung abzubilden. In Abschnitt 3.2.2.4 wurde ein Teil des Modells fokussiert und auf die Testbearbeitung spezifiziert. Dieses Modell soll empirisch getestet werden.

Hypothese 1:

Das in Abschnitt 3.2.2.4 dargestellte Erwartung-Wert-Modell der Motivation zur Testbearbeitung lässt sich durch die empirischen Daten global bestätigen.

Neben einer globalen Prüfung der Modellgeltung sollen einige spezifische Annahmen innerhalb des Modells getestet werden. Nachdem Atkinson (1957, 1964) im Risikowahl-Modell von einer inversen Beziehung zwischen Erwartungs- und Wertkomponente der Leistungsmotivation ausging, hat sich in aktuellen Leistungsmotivationstheorien die Annahme eines positiven Zusammenhangs etabliert (Wigfield & Eccles, 1992; vgl. Abschnitte 3.2.1.4, 3.2.2).

Hypothese 2:

Zwischen der Erwartungs- und der Wertkomponente des Modells besteht ein positiver Zusammenhang.

Die Erwartungskomponente der Leistungsmotivation ist insbesondere für die aktuelle Leistung prädiktiv relevant, während die Wertkomponente vor allem Wahlentscheidungen und langfristige Verhaltensvorhersagen beeinflusst (Abschnitt 3.2.2; vgl. auch Middleton & Toluk, 1999). Zwar ist grundsätzlich auch der Wert positiv mit der Leistung verknüpft, doch verliert sich die Signifikanz dieses Zusammenhangs, wenn Erwartung und Wert zugleich berücksichtigt werden (Schunk et al., 2008).

Hypothese 3:

Für die Vorhersage der Testleistung in einem Leistungstest ohne Aufgabenwahlmöglichkeit ist die Erwartungskomponente wichtiger als die Wertkomponente.

Bereits früh wurde theoretisch auf die Notwendigkeit hingewiesen, bei der Leistungsvorhersage sowohl State- als auch Trait-Aspekte der Leistungsmotivation zu berücksichtigen. Empirisch wurde dieser Anspruch erstmals in Atkinsons Risikowahl-Modell umgesetzt (Atkinson, 1957, 1964; Abschnitt 3.2.1). Auch in dem modernen Erwartung-Wert-Modell von Eccles und Wigfield (2002) sind die situative und die dispositionelle Leistungsmotivation verankert (Abschnitt 3.2.2). Es ist davon auszugehen, dass beide Motivationsaspekte einen statistisch bedeutsamen Anteil an der Leistungsvorhersage haben. In den bestehenden empirischen Studien zur Motivation zur Testbearbeitung in adaptiven Tests wurde hingegen lediglich die State-Motivation berücksichtigt (vgl. Abschnitt 4.2).

Hypothese 4:

Die Modellanpassung des Erwartung-Wert-Modells der Motivation zur Testbearbeitung ist unter Berücksichtigung der State- und der Trait-Leistungsmotivation besser als bei alleiniger Betrachtung des State-Aspekts.

Die Hypothesen zur zweiten Fragestellung werden sowohl auf Basis der theoretischen Grundlagen als auch auf Basis der empirischen Befunde abgeleitet. Aufgrund der theoretischen Annahme, dass in Testsituationen ohne Wahlmöglichkeit insbesondere die Erwartungskomponente die Testleistung beeinflusst (Hypothese 3), gilt für alle nachfolgenden Hypothesen, dass sich mögliche Effekte stärker in der Erwartung als im Wert zeigen sollten (vgl. Wigfield & Eccles, 2002). Der Wert ist für die folgenden Hypothesen darüber hinaus wenig relevant, da er bei fehlender Aufgabenwahlmöglichkeit kaum eine Wirkung auf die Testleistung ausübt. Im Folgenden wird der Einfluss verschiedener Personen- und Testmerkmale (insbesondere FIT im Vergleich zu CAT) auf die Motivation zur Testbearbeitung differenziert untersucht. Es werden daher vor allem spezifische Interaktionshypothesen formuliert. Ein universeller Haupteffekt des Testalgorithmus auf die Motivation zur Testbearbeitung ist theoretisch nicht abzuleiten.

Hypothese 5:
Die mittlere Motivation zur Testbearbeitung im FIT und im CAT unterscheidet sich nicht signifikant.

Im CAT können die Testpersonen ihre tatsächliche Leistung schlechter einschätzen als im FIT (Betz & Weiss, 1976b). Dies liegt vermutlich daran, dass der CAT eine ungewohnte Testsituation darstellt und die Testpersonen üblicherweise nicht wissen, dass die Auswahl der Aufgaben hinsichtlich ihrer Schwierigkeit an die individuelle Leistungsausprägung angepasst wird. Da die Testpersonen ihre Testleistung lediglich in Abhängigkeit von der Anzahl korrekt gelöster Aufgaben wahrnehmen (Tonidandel et al., 2002), entsteht im CAT eine größere Diskrepanz zwischen wahrgenommener Testleistung und tatsächlicher Testleistung. Dies mag sich in einer Diskrepanz zwischen dem Fähigkeitsselbstkonzept und der wahrgenommenen Testleistung widerspiegeln und folglich zu einer Beeinflussung der aktuellen Motivation im Sinne einer Anpassung der Erfolgserwartung führen. Je nach vorherigen Erfahrungen, welche sich beispielsweise im Fähigkeitsselbstkonzept ausdrücken, ist von einer positiven oder negativen Diskrepanz auszugehen (Abschnitte 3.2.2.1 und 3.2.2.4). Im FIT, der die gewohnte Testsituation darstellt und auf dessen Erfahrungen sich das Fähigkeitsselbstkonzept der Testpersonen herausgebildet hat, ist keine Diskrepanz zwischen Fähigkeitsselbstkonzept und wahrgenommener Testleistung zu erwarten.

Hypothese 6:
Der Effekt des Testalgorithmus auf die Motivation zur Testbearbeitung ist vom Fähigkeitsselbstkonzept abhängig.

Hypothese 6a:
Personen mit hohem Fähigkeitsselbstkonzept zeigen im CAT eine niedrigere Motivation zur Testbearbeitung als im FIT.

Hypothese 6b:
Personen mit niedrigem Fähigkeitsselbstkonzept zeigen im CAT eine höhere Motivation zur Testbearbeitung als im FIT.

Die in Hypothese 6 angenommenen, vom Selbstkonzept abhängenden differentiellen Effekte des Testalgorithmus auf die Motivation können die Testfairness und die Validität der Ergebnisse beeinträchtigen. Differentielle Effekte von CAT können möglicherweise vermieden werden, wenn die Testpersonen vor Beginn der Testbearbeitung über den Testalgorithmus aufgeklärt werden (vgl. Ortner & Caspers, in press). Die so geschaffene Transparenz mag bewirken, dass zur Einschätzung der eigenen Testleistung nicht mehr die Quantität der korrekt gelösten Aufgaben verwendet wird, sondern dass auch die Qualität der gelösten Aufgaben in Betracht gezogen wird (vgl. Abschnitte 4.1. und 4.2). Auf diese Weise mag sich die Diskrepanz zwischen wahrgenommener Leistung und Fähigkeitsselbstkonzept verringern, so dass differentielle Effekte auf die Motivation zur Testbearbeitung ausbleiben.

Hypothese 7:
Bei intransparenter Testinstruktion zeigt sich der in Hypothese 6 formulierte Interaktionseffekt.

Hypothese 8:
Bei transparenter Testinstruktion zeigt sich der in Hypothese 6 formulierte Interaktionseffekt nicht.

Eine andere Möglichkeit, differentielle Effekte auf die Motivation zur Testbearbeitung zu unterbinden, könnte darin bestehen, die mittlere Lösungswahrscheinlichkeit des CAT von 50 Prozent auf 70 Prozent zu erhöhen und damit die Diskrepanz zwischen wahrgenommener Testleistung und Fähigkeitsselbstkonzept zu mindern (Andrich, 1995; Betz & Weiss, 1976b; Tonidandel et al., 2002). Denn eine Lösungswahrscheinlichkeit von 70 Prozent entspricht in etwa der durchschnittlichen

Lösungswahrscheinlichkeit im FIT (Lunz & Bergstrom, 1994). Durch diese Manipulation könnte vermieden werden, dass insbesondere hoch leistungsfähige Testpersonen den CAT als „traumatisch“ erleben (Linacre, 2000; vgl. auch Lunz et al., 1994; vgl. Abschnitt 4.2).

Hypothese 9:

Die Motivation zur Testbearbeitung ist im CAT mit geringer Schwierigkeit höher als im CAT mit mittlerer Schwierigkeit.

Hypothese 10:

Bei geringer Schwierigkeit zeigt sich der in Hypothese 6 formulierte Interaktionseffekt nicht.

Da sich frühere Erfahrungen mit Leistungstestsituationen auch in den affektgeprägten Erinnerungen widerspiegeln, die als distale Größe aufzufassen und im Sinne eines Leistungsmotivs zu interpretieren sind, mögen die Effekte der Testmerkmale und des Fähigkeitsselbstkonzepts auf die Motivation zur Testbearbeitung je nach Ausprägung des Leistungsmotivs unterschiedlich ausfallen. Wegen der wichtigen Rolle, die die besuchte Schulart für die Ausbildung des Fähigkeitsselbstkonzepts spielt, mögen sich die Zusammenhänge zwischen den untersuchten Testmerkmalen, dem Fähigkeitsselbstkonzept und der Motivation zur Testbearbeitung zudem in Abhängigkeit von der Schulart unterschiedlich gestalten. Daher werden die Hypothesen zur zweiten Fragestellung nicht nur anhand der Gesamtstichprobe analysiert, sondern auch explorativ anhand von Substichproben nach Leistungsmotivausprägung und nach Schulart.

Neben der allgemeinen Prüfung des auf die Testsituation zugeschnittenen Erwartung-Wert-Modells der Leistungsmotivation leisten die Hypothesentests einen Beitrag dazu, bislang nur theoretisch formulierte Vorschläge und Ideen zur Optimierung computerisierter adaptiver Testverfahren empirisch zu prüfen. Es ist zu erwarten, dass diese Arbeit sowohl theoretisch als auch empirisch einen wichtigen Beitrag zum näheren Verständnis der Motivation zur Testbearbeitung in computerisierten adaptiven und nicht-adaptiven Leistungstests erbringt.

# 6 Methode

Die Studie wurde mit Schülerinnen und Schülern der neunten Jahrgangsstufe durchgeführt. Diese Stichprobe wird in Abschnitt 6.1 beschrieben. Es wurden Tests zur mathematischen Kompetenz und Fragebögen zur Leistungsmotivation vorgegeben. Außerdem wurden wenige personenbezogene Fragen gestellt. Die Erhebungsinstrumente werden in Abschnitt 6.2 vorgestellt. Die Studie basiert auf einem experimentellen Versuchsplan, der in Abschnitt 6.3 dargestellt wird. Abschnitt 6.4 dient der Erläuterung der Versuchsdurchführung, die sich über zwei Schulstunden erstreckte. Der Methodenteil wird mit Beschreibungen zur statistischen Analyse der Daten in Abschnitt 6.5 abgeschlossen.

## 6.1 Stichprobe

Aufgrund der theoretischen Vorüberlegungen und der Erkenntnisse aus bestehenden empirischen Studien zur Motivation zur Testbearbeitung wurde im Hinblick auf die Stichprobe Folgendes angestrebt: Erstens sollte die Stichprobe hinreichend groß sein, um die Interaktionshypothesen testen zu können und die Effekte auf die Motivation zur Testbearbeitung abbilden und mit akzeptabler Teststärke statistisch gegen den Zufall absichern zu können. Im Rahmen einer Vorstudie wurde eine anzustrebende Stichprobengröße von $N_{opt} = 800$ bestimmt.

Zweitens sollten Schülerinnen und Schüler der neunten Jahrgangsstufe untersucht werden. Der theoretische Hintergrund für diese Entscheidung ist, dass in der Adoleszenz bereits differenzierte Kompetenzüberzeugungen bestehen und dass die Jugendlichen bereits über umfassende Erfahrung im Bereich von Mathematiktests in Form von Klassenarbeiten verfügen (Eccles & Wigfield, 1995). Es ist davon auszugehen, dass diese mathematikbezogenen Erfahrungen die motivationalen Überzeugungen und Dispositionen und die Erwartungshaltung in Bezug auf eine erfolgreiche Bearbeitung des Mathematiktests prägen (vgl. Ausführungen zum Erwartung-Wert-Modell der Motivation zur Testbearbeitung, Abschnitt 3.2.2.4). Der praktische Hintergrund dieser Entscheidung ist, dass eine Nähe zu PISA als Beispiel für groß angelegte Vergleichsstudien angestrebt wurde, da eine Einführung computerisierten adaptiven Testens in solchen Studien besonders effektiv wäre und auf internationaler Ebene diskutiert wird. Bei PISA werden 15-Jährige untersucht; in Deutschland wurden für vertiefende nationale Analysen zusätzlich vollständige neunte Klassen erhoben.

Drittens sollte die Stichprobe eine breit gestreute Leistungsverteilung gewährleisten, da die Motivation zur Testbearbeitung eng mit individuellen Kompetenzüberzeugungen verknüpft ist, die in Abhängigkeit von der tatsächlichen Leistungsausprägung der Testpersonen und von früher gemachten Testerfahrungen, zum Beispiel in Klassenarbeiten, variieren. In nahezu allen empirischen Studien, die sich mit der Motivation zur Testbearbeitung beschäftigt haben, wurden jedoch studentische oder gymnasiale Stichproben verwendet (vgl. Abschnitt 4.3). Dies führt von vorneherein zu einer Einschränkung des Geltungsbereichs der Ergebnisse und zu einer Einschränkung der Generalisierbarkeit auf allgemeine Grundgesamtheiten. Vor dem Anwendungskontext groß angelegter Vergleichsstudien wie PISA ist angezeigt, eine leistungsheterogene Stichprobe zu untersuchen. Mögliche systematische Verzerrungen der Testergebnisse, beispielsweise durch

differentielle Effekte von Fähigkeitsselbstkonzept und Testalgorithmus auf die Motivation zur Testbearbeitung, und eine daraus folgende mögliche Beeinträchtigung der Testfairness können andernfalls nur eingeschränkt aufgedeckt werden. Im Rahmen der vorliegenden Studie sollte die breite Leistungsverteilung durch eine Untersuchung von Neuntklässlerinnen und Neuntklässlern aller Schularten gewährleistet werden. Da die neunte Jahrgangsstufe in Deutschland noch zur Pflichtschulzeit zählt, war zu erwarten, dass eine Heterogenität der mathematischen Kompetenz in dieser Stichprobe gut zu realisieren ist.

Die vorliegende Stichprobe erfüllt alle drei Anforderungen. Unter Orientierung an der Fachserie 11, Reihe 1 für allgemeinbildende Schulen des Statistischen Bundesamtes für das Schuljahr 2007/2008 (Statistisches Bundesamt, 2010, Tabelle 3.4) wurde die prozentuale Verteilung der Neuntklässlerinnen und Neuntklässler in Schleswig-Holstein auf die Schularten Hauptschule, Integrierte Gesamtschule, Realschule und Gymnasium berechnet und als Populationsverteilung verwendet. Da die Umsetzung der Schulreform in Schleswig-Holstein zum Zeitpunkt der Erhebung uneinheitlich weit fortgeschritten war, beruht die Zuteilung der Schulen ausschließlich auf diesen traditionellen Schulart-Kategorien. Die Rekrutierung der Schulen für die Stichprobe geschah auf zweierlei Wegen: Zum einen erfolgte pro Schulart eine Zufallsziehung von Schulen und Klassen durch das *Data Processing and Research Center* (DPC) in Hamburg. Zum anderen wurden bestehende Kontakte des Leibniz-Instituts für die Pädagogik der Naturwissenschaften und Mathematik (IPN) in Kiel genutzt, um Schulen anzuschreiben, die an dem Programm SINUS-Transfer teilnehmen (SINUS = „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“). Von insgesamt 54 schriftlich angefragten Schulen sagten 20 Schulen mit insgesamt 34 neunten Klassen eine Teilnahme an der Studie zu (Teilnahmequote: 37 %; an mehrzügigen Schulen nahmen nach Möglichkeit zwei zufällig gezogene neunte Klassen an der Studie teil). Die 20 Schulen bestehen aus sieben Hauptschulen, einer Integrierten Gesamtschule, sechs Realschulen und sechs Gymnasien. Die Verteilung der gezogenen Schülerinnen und Schüler auf die Schularten kommt der Populationsverteilung nahe (Abbildung 6.1).

Die Genehmigung der Studie durch das Ministerium für Bildung und Frauen des Landes Schleswig-Holstein sowie Elterngenehmigungen für die Erfassung der personenbezogenen Angaben wurden vor Durchführung der Studie eingeholt. Die Teilnahme an den Mathematiktests und den Fragebögen war für die Jugendlichen verpflichtend. Die Beantwortung der personenbezogenen Fragen war freiwillig und deren Auswertung an das Vorliegen der Elterngenehmigung gebunden. Die Datenerhebung fand im September und Oktober 2009 und damit zu Beginn des Schuljahres statt. Nach Abschluss der gesamten Datenerhebung wurde den jeweiligen Klassenlehrerinnen und Klassenlehrern bei Interesse eine Rückmeldung über die mathematische Kompetenz und die Leistungsmotivation ihrer Klasse im Vergleich zu den anderen teilnehmenden Klassen derselben Schulart zugeschickt (vgl. Abschnitt 6.2). Alle beteiligten Lehrkräfte machten von diesem Angebot Gebrauch.

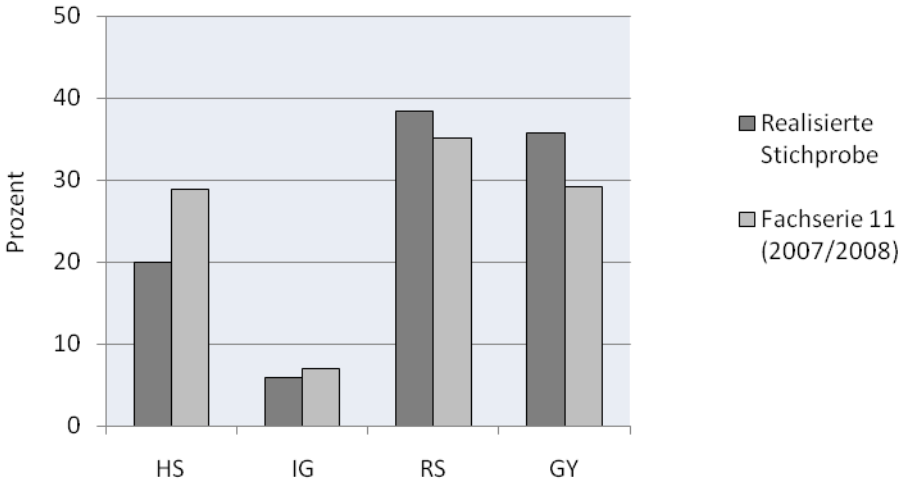



*Abbildung 6.1*: Prozentuale Verteilung der Schülerinnen und Schüler auf die Schularten in Schleswig-Holstein und in der Stichprobe (HS = Hauptschule, IG = Integrierte Gesamtschule, RS = Realschule, GY = Gymnasium).

Die gezogene Stichprobe bestand aus insgesamt *N* = 796 Schülerinnen und Schüler der neunten Jahrgangsstufe. Krankheitsbedingt nahmen 34 gezogene Personen nicht an der Studie teil (4 %). Damit ergab sich im Test eine durchschnittliche Klassengröße von 22 Personen (*SD* = 3). Von den 762 getesteten Schülerinnen und Schülern mussten die Daten von 40 Personen (5 %) aufgrund technischer Probleme bei der Durchführung des Computerprogramms von der Analyse ausgeschlossen werden. Die technischen Probleme bestanden in den meisten Fällen darin, dass das Weiterblättern auf die nächste Seite nicht möglich war. In wenigen Fällen führte unsachgemäßes Verhalten der Testperson zu einem Absturz des Computer-Programms. Ein Netbook musste wegen eines produktionsbedingten Hardware-Schadens von weiteren Testungen ausgeschlossen werden. Für 13 Personen (2 %) lagen keine Elterngenehmigungen vor, so dass auch die Daten dieser Jugendlichen eliminiert wurden. Von den verbliebenen 709 Personen wurden schließlich sechs Personen (1 %) wegen mangelnder Deutschkenntnisse von den weiteren Analysen ausgeschlossen. Unter dieses Ausschlusskriterium fielen Jugendliche, deren Muttersprache nicht Deutsch ist und die angaben, Schwierigkeiten beim sprachlichen Verständnis der Texte der Studie gehabt zu haben sowie generell kaum einen Text auf Deutsch zu verstehen.

Die endgültige Stichprobe besteht aus *N* = 703 Schülerinnen und Schülern der neunten Jahrgangsstufe (davon 48 % weiblich). Das Alter der Testpersonen beträgt 13 bis 18 Jahre (*M* = 14.54, *SD* = 0.73). Für 596 Personen in der Stichprobe (85 %) ist die Muttersprache Deutsch.

## 6.2 Erhebungsinstrumente

Die eingesetzten Erhebungsinstrumente lassen sich in drei Gruppen unterteilen: Zum einen wurden Tests zur Messung der mathematischen Kompetenz vorgegeben (Abschnitt 6.2.1). Zum anderen wurden mehrere Fragebögen zur Erfassung der State- und der Trait-Aspekte der Leistungsmotivation verwendet (Abschnitt 6.2.2). Schließlich wurden den Schülerinnen und Schülern einige

personenbezogene Fragen gestellt (Abschnitt 6.2.3). Alle Erhebungen wurden vollständig computerisiert durchgeführt. Dazu wurde ein Klassensatz (30 Stück) 10‘‘-Netbooks von Lenovo (IdeaPad S10e) genutzt. Die Netbooks verfügen über eine Bildschirm-Auflösung von 1024x600 Bildpunkten, eine Rechengeschwindigkeit von 1.6 GHz, einen 1-GB-RAM-Speicher und eine 160-GB-Festplatte.

### 6.2.1 Tests zur mathematischen Kompetenz

Zur Erfassung der mathematischen Kompetenz wurden ein adaptiver (Abschnitt 6.2.1.1) und ein nicht-adaptiver Kompetenztest (Abschnitt 6.2.1.2) verwendet. Grundlage der Kompetenztests bildeten Aufgaben, die im Zuge der Normierung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss entwickelt worden waren (Prenzel & Blum, 2007; Kalibrierungsstichprobe: $N$ = 9577 Neuntklässlerinnen und Neuntklässler in Deutschland; davon 49 % weiblich; Alter: $M$ = 15.7, $SD$ = 0.57). Für die Tests wurden wegen der notwendigen unmittelbaren computerisierten Auswertung lediglich Aufgaben mit einfachem Multiple-Choice-Antwortformat ausgewählt. Zudem wurde sichergestellt, dass sich der Aufgabenstamm gut auf einer Bildschirmseite der verwendeten Netbooks darstellen ließ. Die meisten Bildungsstandards-Aufgaben enthielten ursprünglich pro Aufgabenstamm mehrere Aufgaben (so genannte Aufgabeneinheiten). In der vorliegenden Studie wurde pro Aufgabeneinheit jeweils nur eine Aufgabe verwendet. So wird vermieden, dass dieselbe Person mehrmals denselben Aufgabenstamm (wenn auch mit unterschiedlichen Aufgaben) vorgegeben bekommen könnte, was unerwünschte Effekte wie zum Beispiel Reaktanz oder Langeweile hervorrufen könnte.

Insgesamt erfüllten 112 Aufgaben diese Bedingungen. Von diesen 112 Aufgaben wurden zwei Aufgaben als Aufgabenbeispiele genutzt, die vor Beginn der eigentlichen Kompetenztests präsentiert wurden. Die übrigen 110 Aufgaben standen für die Zusammenstellung der unterschiedlichen Tests zur Verfügung. Jeweils zehn Aufgaben bildeten einen adaptiven und einen nicht-adaptiven Kompetenztest (*fixed test length*, vgl. Abschnitt 2.2). Die Aufgaben pro Test waren nicht für jede Testperson identisch. Die Gründe hierfür und der genaue Aufbau der Tests werden im Folgenden näher erläutert.

#### *6.2.1.1 Adaptiver Test zur mathematischen Kompetenz*

Um den adaptiven Kompetenztest zu beschreiben, werden die in Abschnitt 2.2 formulierten Bestimmungsstücke spezifiziert: Als Item-Response-Modell wurde das Rasch-Modell gewählt. Dieses Modell wurde auch bei der Skalierung der Bildungsstandards-Aufgaben verwendet. Der Aufgabenpool bestand aus insgesamt 100 Aufgaben, da 10 der oben genannten 110 Aufgaben für den nicht-adaptiven Test verwendet wurden (vgl. auch Abschnitt 6.3). Aus diesem Itempool wurden nach einer für alle Testpersonen identischen Anfangsaufgabe auf Basis des individuellen Antwortverhaltens neun Aufgaben ausgewählt und vorgegeben. Der Itempool erfüllte also die Faustregel, etwa acht- bis zwölfmal so viele Aufgaben zu beinhalten wie Aufgaben vorgegeben werden (de Ayala, 2009; vgl. auch Weiss & Kingsbury, 1984). Die Schwierigkeiten der Aufgaben streuten zwischen -4.22 und 3.48 Logits ($M$ = -0.08; $SD$ = 1.49).

Die Anfangsaufgabe war eine Aufgabe mittlerer Schwierigkeit ($b$ = 0.01; vgl. Abschnitt 7.1 zu Einzelheiten der Skalierung). Die Auswahl der folgenden Aufgaben geschah auf Basis des Kriteriums maximaler Information. Da in der vorliegenden Studie auch die Testschwierigkeit variiert wurde (vgl. Abschnitt 6.3), wurden zwei verschiedene CAT-Versionen erstellt, die sich lediglich in der Aufgabenauswahl unterschieden: Neben dem herkömmlichen CAT mit einer Aufgabenauswahl anhand der maximalen Information zum vorläufigen Personenparameterschätzer („mittlere Testschwierigkeit“) gab es einen CAT, dessen Algorithmus zur Aufgabenauswahl manipuliert wurde („geringe Testschwierigkeit“). Die Manipulation erfolgte durch eine Verminderung des vorläufigen Personenparameterschätzers um einen Betrag von 0.847 Logits, bevor die nächste Aufgabe ausgewählt wurde. Auf diese Weise wurde eine geringere Kompetenz der Testpersonen simuliert, was folglich zu einer Auswahl einfacherer Aufgaben führte. Die Verminderung um 0.847 Logits ergibt sich aus folgendem Grund: Aus der in Abschnitt 2.1 vorgestellten logistischen Funktion des Rasch-Modells ergibt sich, dass die Lösungswahrscheinlichkeit 50 Prozent beträgt, wenn sich Personenparameter und Aufgabenschwierigkeit entsprechen. Um eine höhere Lösungswahrscheinlichkeit als 50 Prozent zu realisieren, muss der Personenparameter die Aufgabenschwierigkeit übertreffen. Um welchen Betrag sich die beiden Parameter unterscheiden müssen, um eine Lösungswahrscheinlichkeit von 70 Prozent zu erreichen, kann über die Formel einfach ausgerechnet werden:

$$P(x_{\mathrm{vi}} = 1 \,|\theta_{\mathrm{v}}, \mathrm{b_i}) = \frac{\mathrm{e}^{(\theta_{\mathrm{v}} - \mathrm{b_i})}}{1 + \mathrm{e}^{(\theta_{\mathrm{v}} - b_{\mathrm{i}})}} = 0.7$$

Löst man diese Formel nach $\mathrm{e}^{(\theta_{\mathrm{v}} - \mathrm{b_i})}$ auf und bildet den natürlichen Logarithmus, ergibt sich:

$$\theta_{\mathrm{v}} - b_{\mathrm{i}} = \ln\frac{0.7}{0.3} = 0.847$$

Der vorläufige Personenparameter $\hat{\theta}_{\mathrm{v}}$ muss den Aufgabenschwierigkeitsparameter $b_{\mathrm{i}}$ also um einen Betrag von 0.847 übertreffen, damit Person *v* Aufgabe *i* mit einer 70-prozentigen Wahrscheinlichkeit lösen kann. Andersherum ausgedrückt müssen die Schwierigkeiten der ausgewählten Aufgaben um 0.847 kleiner sein als der vorläufige Personenparameterschätzer, was in der einfachen Testbedingung dadurch gewährleistet wurde, dass der vorläufige Personenparameterschätzer für die Aufgabenauswahl um diesen Betrag vermindert wurde. Die Manipulation erfolgte erst nach Ausspeicherung des tatsächlichen vorläufigen Personenparameterschätzers, da lediglich die Aufgabenauswahl beeinflusst werden sollte. Nach Vorgabe der letzten Aufgabe entfiel die Verminderung des Personenparameterschätzers.

In beiden CAT-Versionen wurden zum Zwecke der an das individuelle Kompetenzniveau angepassten Aufgabenauswahl vorläufige und endgültige Personenparameter in Form von EAPs geschätzt (vgl. Abschnitt 2.2). Als Abbruchkriterium wurde eine fixe Testlänge von zehn Aufgaben definiert. Bereits bei dieser Testlänge liegen recht präzise Personenparameterschätzer vor (Weiss, 1982).

#### *6.2.1.2 Nicht-adaptiver Test zur mathematischen Kompetenz*

Der nicht-adaptive Test bestand ebenfalls aus zehn Aufgaben, die jedoch vorab festgelegt wurden. Um den Einfluss bestimmter Aufgabeninhalte auf die Testleistung einzuschränken und Effekte der Darbietungsposition einer Aufgabe auf die beobachtete Lösungshäufigkeit zu minimieren, wurden sechs verschiedene FIT-Versionen erstellt (Tabelle 6.1). Jede Testperson wurde zufällig einer FIT-

Version zugewiesen. Drei der sechs Versionen hatten eine mittlere Aufgabenschwierigkeit nahe $b = 0$ ($b_1 = 0.06$, $b_2 = 0.03$, $b_3 = 0.04$) und repräsentierten die Testbedingung mittlerer Schwierigkeit (vgl. Erläuterungen zum Design in Abschnitt 6.3). Die drei anderen Versionen hatten eine mittlere Aufgabenschwierigkeit nahe $b = -0.5$ ($b_4 = -0.50$, $b_5 = -0.53$, $b_6 = -0.51$) und repräsentierten die Testbedingung geringer Schwierigkeit. Jeder FIT bestand aus zwei Clustern von jeweils fünf Aufgaben (Tabelle 6.1). Für jede der beiden Schwierigkeitsbedingungen gab es drei Aufgabencluster, die sich so auf die FIT-Versionen verteilten, dass jedes Cluster einmal an erster Position und einmal an zweiter Position auftrat. Die Reihenfolge der Aufgaben innerhalb der Cluster wurde nicht variiert; die Aufgabenschwierigkeiten stiegen innerhalb eines Clusters von Aufgabe zu Aufgabe an.

*Tabelle 6.1:* Verteilung der Aufgabencluster A bis F auf die FIT-Versionen (jedes Cluster enthält fünf Aufgaben).

| | FIT-Version | | | | | |
|---|---|---|---|---|---|---|
| | Mittelschwierig | | | Einfach | | |
| Position | 1 | 2 | 3 | 1 | 2 | 3 |
| 1 | A | B | C | D | E | F |
| 2 | B | C | A | E | F | D |

Die computerisierten adaptiven und nicht-adaptiven Kompetenztests wurden in Zusammenarbeit mit der Arbeitsgruppe *Technology Based Assessment* des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) in Frankfurt/Main erstellt. Bei der Programmierung wurde auf die Open-Source-Plattform *Testing Assisté Par Ordinateur* (TAO; http://www.tao.lu) zurückgegriffen. TAO ist eine von der Universität Luxemburg entwickelte Plattform für computerisierte Leistungstests, die auch im Rahmen von PISA (zur Erfassung der Kompetenz zum Lesen von elektronischen Texten; vgl. Martens, Goldhammer, Rölke, Scharaf & Upsing, 2008; siehe auch http://www.tao/lu/) und PIAAC (*Programme for the International Assessment of Adult Competencies*; OECD, 2010) eingesetzt wird. Um eine gute Lesbarkeit sicherzustellen, wurden alle Texte in serifenloser, hinreichend großer Schrift dargestellt (Arial, 10pt; vgl. Georgiadou et al., 2006). Um ein versehentliches oder absichtliches Verlassen des Programms durch die Testperson zu unterbinden, geschah die Präsentation der Studie im so genannten Kiosk-Modus, das heißt, ohne obere oder untere Browser-Menüzeile. Gängige Tastenkombinationen wie Strg+Alt+Entf wurden deaktiviert. Das Herunterfahren des Netbooks oder das Aktivieren des Ruhezustands über das Drücken des Einschaltknopfes waren nicht möglich.

### 6.2.2 Fragebögen zur Leistungsmotivation

Für ein gründliches Verständnis der während einer Testsituation ablaufenden motivationalen Prozesse erscheint es notwendig, die Motivation differenziert und wiederholt zu erfassen (vgl. Abschnitt 3.2.2). Daher wurden in der vorliegenden Studie verschiedene Motivationsfragebögen eingesetzt, von denen einige die State- und andere die Trait-Aspekte der Leistungsmotivation messen.

Die aktuelle Motivation während und nach der Testbearbeitung (State-Motivation) wurde zum einen mit dem *On-Line Motivation Questionnaire* (OMQ; Boekaerts, 2002) erfasst, der in Teilen auch in den nationalen Ergänzungen von PISA 2000 bis 2006 in Deutschland eingesetzt wurde (Frey, Taskinen et al., 2009; Kunter et al., 2002; Ramm et al., 2006; vgl. auch Baumert & Demmrich, 2001). Der OMQ bildet die situationsspezifische Leistungsmotivation speziell in Kompetenztestsituationen ab. Boekaerts entwickelte diesen Fragebogen vor dem Hintergrund, dass im Verlauf der Testbearbeitung ununterbrochen ein Vergleichsprozess zwischen den aktuellen situationsspezifischen Anforderungen und den persönlichen Ressourcen stattfindet, diesen Anforderungen zu genügen (Crombach, Boekaerts & Voeten, 2003). Der OMQ besteht aus zwei Teilen. Der eine wird während der Testung, der andere direkt im Anschluss an die Testung vorgegeben. Jeder Fragebogenteil enthält mehrere voneinander unabhängige Skalen (Crombach et al., 2003), von denen im Rahmen der vorliegenden Studie folgende eingesetzt wurden: Während der Testung wurden die *Erwartungskomponente* (Erfolgserwartung und Selbstwirksamkeitserwartung) und die *Wertkomponente* (Nützlichkeit, Wichtigkeit und Interesse) erfasst. Nach der Testung wurde erneut die *Erwartungskomponente* erfragt. Obwohl die nach dem Test eingesetzte Skala lediglich aus zwei Fragen besteht, hat sich die Skala für Analysen auf latenter Ebene bewährt (Crombach et al., 2003). Alle Fragen wurden auf 4-stufigen Likert-Skalen beantwortet, deren Pole mit 1 = *gar nicht/gar keine* und 4 = *sehr/sehr viel* bezeichnet waren. Für die Skalen wurden Mittelwerte gebildet. Die in der vorliegenden Studie beobachteten Reliabilitäten der Skalen (Cronbachs Alpha) sind zufriedenstellend bis gut (Tabelle 6.2). Für die Validität des OMQ liegen zahlreiche empirische Befunde vor (vgl. Übersicht in Boekaerts, 2002; vgl. auch Crombach et al., 2003; Karabenick et al., 2007; Röll, 1994).

Zum anderen wurde die Wertkomponente der aktuellen Motivation zur Testbearbeitung nach der Testung über die *Student Opinion Scale* (SOS; Sundre, 2007) erfasst. Ein zusätzliches Instrument zur Erfassung des Werts erschien notwendig, da die Ausprägung der Wertkomponente nach Beendigung der Testung im OMQ nicht als eigenständige Skala berücksichtigt wird. Die SOS wurde auf der theoretischen Basis des Erwartung-Wert-Modells nach Eccles et al. (1983) entwickelt. Im Rahmen der vorliegenden Studie wurde die Skala *Wichtigkeit* (als Teil der Wertkomponente) genutzt. Die fünf Fragen dieser Skala werden auf 5-stufigen Likert-Skalen von 1 = *stimme gar nicht zu* bis 5 = *stimme voll zu* beantwortet. Für die Auswertung wurde der Skalenmittelwert gebildet. Die in der vorliegenden Studie beobachtete Reliabilität der Skala (Cronbachs Alpha) ist zufriedenstellend (Tabelle 6.2). Die Validität des SOS wurde in mehreren Studien geprüft und bestätigt (z. B. Sundre, 2007; Sundre & Finney, 2002; Sundre & Moore, 2002). Die SOS eignet sich besonders für den Einsatz in Low-Stakes-Tests wie denen der vorliegenden Studie.

Um die subjektive Interpretation der Testerfahrung zu erfassen, wurde eine selbst formulierte Frage eingesetzt, die sich auf die wahrgenommene Testleistung bezieht (Tabelle 6.2). Diese Frage wurde auf einer 8-stufigen Likert-Skala mit den Polen 1 = *überhaupt nicht zufrieden* bis 8 = *sehr zufrieden* beantwortet. Die Formulierung der Frage zielt darauf ab, eine Einschätzung der wahrgenommenen Leistung im Vergleich zur subjektiv erwarteten Leistung zu erhalten, nicht eine Einschätzung der objektiven Höhe der Leistung im Vergleich zur Leistung anderer Personen oder zu einem allgemeinen Maßstab.

*Tabelle 6.2:* Überblick über die verwendeten Fragebogenskalen (vgl. Abschnitt 6.4).

| Skala | Beispielfrage/-aussage | $N_{Fragen}$ | *M* | *SD* | α |
|---|---|---|---|---|---|
| Erwartung während des Tests (OMQ) | Was meinst du, wie gut du bei den folgenden Mathematikaufgaben abschneiden wirst? | 6 | $t_1$: 2.49<br>$t_2$: 2.43 | $t_1$: 0.55<br>$t_2$: 0.64 | $t_1$: .84[3]<br>$t_2$: .88[3] |
| Wert während des Tests (OMQ) | Wie wichtig findest du es, in diesem Test gut abzuschneiden? | 5 | $t_1$: 2.71<br>$t_2$: 2.48 | $t_1$: 0.64<br>$t_2$: 0.75 | $t_1$: .78<br>$t_2$: .86 |
| Erwartung nach dem Test (OMQ) | Wie gut hast du wohl in diesem Test abgeschnitten? | 2 | $t_1$: 2.35<br>$t_2$: 2.28 | $t_1$: 0.66<br>$t_2$: 0.72 | $t_1$: .77<br>$t_2$: .83 |
| Wert nach dem Test (SOS) | In diesem Test gut abzuschneiden, war wichtig für mich. | 5 | $t_1$: 3.15<br>$t_2$: 3.06 | $t_1$: 0.80<br>$t_2$: 0.81 | $t_1$: .77[4]<br>$t_2$: .76[4] |
| Wahrgenommene Leistung während des Tests | Wie zufrieden bist du mit deinen Antworten auf die Mathematikaufgaben? | 1 | $t_1$: 4.83<br>$t_2$: 4.82 | $t_1$: 1.79<br>$t_2$: 1.95 | - |
| Wahrgenommene Leistung nach dem Test | Wie zufrieden bist du mit deinen Antworten auf die Mathematikaufgaben? | 1 | $t_1$: 4.54<br>$t_2$: 4.50 | $t_1$: 1.91<br>$t_2$: 1.98 | - |
| Fähigkeitsselbst-konzept | Mathematik liegt mir nicht besonders. | 5 | 2.82 | 0.73 | .85 |
| Selbstwirksam-keitserwartung | Wie sicher glaubst du, folgende Mathematikaufgaben lösen zu können? Ausrechnen, wie viel billiger ein Fernseher bei 30 % Rabatt wäre | 4 | 2.85 | 0.58 | .65 |
| Hoffnung auf Erfolg | Mich reizen Situationen, in denen ich meine Fähigkeiten testen kann. | 5 | 2.77 | 0.64 | .81 |

*Anmerkungen.* t1: 1. Teststunde, t2: 2. Teststunde; OMQ: On-Line Motivation Questionnaire; SOS: Student Opinion Scale.

Außerdem wurden das *Fähigkeitsselbstkonzept* und die *Selbstwirksamkeitserwartung* in Mathematik erhoben, da diese motivationalen Überzeugungen für die Erwartungshaltung, mit der eine Person in eine Testsituation hinein geht, eine wesentliche Rolle spielen (vgl. Abschnitt 3.2.2). In dem Fähigkeitsselbstkonzept spiegeln sich die Leistungserfahrungen wider, die eine Person im Verlauf ihres bisherigen Lebens gemacht hat. Es steht in engem Zusammenhang zu der Erwartungskomponente der aktuellen Motivation. Das mathematische Fähigkeitsselbstkonzept wurde in Anlehnung an Jopt (1978) und Jerusalem (1984) sowie Möller und Köller (2001; vgl. auch Köller et al., 2000; Köller & Möller, 1995; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005; Möller & Köller, 2000) über Aussagen erfasst, die auf 4-stufigen Likert-Skalen zu beantworten waren (von 1 = *trifft voll und ganz zu* bis 4 = *trifft überhaupt nicht zu*). Die in der vorliegenden Studie beobachtete Reliabilität (Cronbachs Alpha) der Selbstkonzept-Skala ist als gut zu bezeichnen (Tabelle 6.2). Die Selbstwirksamkeitserwartung ist tätigkeitsspezifischer definiert als das Fähigkeitsselbstkonzept. Sie bezieht sich auf die Einschätzung, ein konkretes leistungsbezogenes Verhalten ausführen zu können (Möller, 2008). Die Selbstwirksamkeitserwartung wurde über eine verkürzte Version der Skala erfasst, die bei PISA 2003 eingesetzt wurde (Ramm et al., 2006). Das Antwortformat bestand aus

[3] Eine Frage der Erwartungsskala wurde aufgrund schlechter Kennwerte für beide Teststunden eliminiert.
[4] Eine Frage der Wertskala wurde aufgrund schlechter Kennwerte für beide Teststunden eliminiert.

einer 4-stufigen Likert-Skala mit den Polen 1 = *sehr sicher* und 4 = *gar nicht sicher*. Die beobachtete Reliabilität ist als eher niedrig zu bewerten (Tabelle 6.2), was vermutlich der Kürze der Skala anzulasten ist.

Zur Erfassung der Trait-Motivation wurde die Kurzform der revidierten *Achievement Motive Scale* eingesetzt (AMS-R; Lang & Fries, 2006; Originalversion: Gjesme & Nygard, 1970, zitiert nach Dahme, Jungnickel & Rathje, 1993, S. 257). Dieses etablierte und viel verwendete Instrument eignet sich, um die Leistungsmotiv-Komponente *Hoffnung auf Erfolg* in Form eines selbstattribuierten Konstrukts zu messen (vgl. Abschnitt 3). Die Fragen waren auf 4-stufigen Likert-Skalen von 1 = *trifft gar nicht auf mich zu* bis 4 = *trifft völlig auf mich zu* zu beantworten. Die Verwendung eines selbstattribuierten Maßes erscheint im Rahmen der vorliegenden Arbeit als angemessen, weil es das Leistungsverhalten in spezifischen Situationen unmittelbar vorhersagen kann (vgl. McClelland et al., 1989; Spangler, 1992; vgl. Abschnitt 3.2). Die in der vorliegenden Studie beobachtete Reliabilität der Skala ist gut (Cronbachs Alpha; Tabelle 6.2). Lang und Fries (2006) führen mehrere Belege für die Validität der Skala an. Da die Trait-Motivation per definitionem nur träge veränderlich ist, ist im Gegensatz zur State-Motivation keine wiederholte Messung notwendig.

Zur Bildung der Skalenmittelwerte und weitere Analysen wurden die Antworten auf alle Fragebögen im Sinne des jeweiligen Merkmals kodiert.

### 6.2.3 Personenbezogene Angaben

Um die Stichprobe angemessen beschreiben zu können, wurden einige wenige personenbezogene Fragen gestellt. Die Fragen bezogen sich auf das Alter, das Geschlecht der Testperson und die momentan besuchte Schulart. Abschließend wurden Angaben zu den Deutschkenntnissen erbeten (Muttersprache Deutsch; sprachliches Verständnis der Texte in der Studie), um Testpersonen a posteriori von der Auswertung ausschließen zu können, deren mangelnde Deutschkenntnisse eine sinnvolle Bearbeitung der Testaufgaben und der Fragebögen unmöglich machten (vgl. Abschnitt 6.1). Für die Erfassung und die Analyse der personenbezogenen Daten wurden vorab schriftliche Genehmigungen der Eltern eingeholt.

## 6.3 Design

Es wurde ein dreifaktorielles, vollständig gekreuztes Experiment durchgeführt (2x2x2-Design). Die erste unabhängige Variable (UV1) ist der 2-stufige Innersubjektfaktor *Testalgorithmus* (CAT bzw. FIT; vgl. Abschnitt 6.2.1). Diese Variable wurde innerhalb der Testpersonen variiert, um eine ausreichend hohe Anzahl an Personen in den einzelnen Versuchsbedingungen zu gewährleisten. Die Reihenfolge der Vorgabe von CAT und FIT wurde über die Testpersonen hinweg counterbalanciert. So wird erreicht, dass der Mittelwert möglicher Positionseffekte auf die abhängige Variable in allen Versuchsbedingungen in gleicher Weise besteht.

Die zweite unabhängige Variable stellt die *Transparenz der Testinstruktion* (UV2) dar und wurde als Zwischensubjektfaktor 2-stufig variiert. Die Testinstruktion war entweder ohne Erläuterung

des Testalgorithmus formuliert (intransparente Testinstruktion) oder mit Erläuterung des Testalgorithmus (transparente Testinstruktion). In der Intransparenzbedingung wurden keine näheren Angaben über den Testalgorithmus gemacht, der dem Mathematiktest zugrunde lag. In der Transparenzbedingung mit CAT wurde vor Beginn des Mathematiktests darüber aufgeklärt, dass die Aufgaben von ihrer Schwierigkeit her zu der mathematischen Fähigkeit der Testperson passen, dass nach einer richtigen/falschen Antwort eine schwierigere/einfachere Aufgabe folgt und dass so vermieden wird, dass die Testperson Aufgaben bearbeiten muss, die für sie persönlich viel zu einfach oder viel zu schwierig sind. Die Transparenzbedingung mit FIT enthielt vor Beginn des Mathematiktests die Information, dass alle Testpersonen unabhängig von ihrer mathematischen Fähigkeit dieselben Aufgaben bekommen und dass dies dazu führen kann, dass einige Aufgaben für die Testperson viel zu einfach oder viel zu schwierig sind. Alle Testpersonen wurden gebeten, die Aufgaben konzentriert zu bearbeiten. Um nach Abschluss der Studie feststellen zu können, ob die unterschiedliche Formulierung der Testinstruktion in Zusammenhang mit der bearbeiteten Testversion (FIT/CAT) von den Testpersonen rezipiert wurde, wurde nach dem Mathematiktest eine Frage zur Manipulationskontrolle gestellt. Sie erfragte, wie die Aufgaben im Mathematiktest ausgewählt wurden, und enthielt die nominalen Antwortkategorien *Die Aufgaben wurden für mich persönlich nach meiner mathematischen Fähigkeit ausgewählt*, *Die Aufgaben wurden vorab festgelegt und waren für alle Personen gleich* und *Das wurde nicht erklärt*.

Die dritte unabhängige Variable ist die *Testschwierigkeit* (UV3). Sie wurde als Zwischensubjektfaktor 2-stufig variiert, um eine mittlere erwartete Lösungswahrscheinlichkeit des Tests von entweder etwa 50 Prozent (mittlere Testschwierigkeit) oder etwa 70 Prozent zu realisieren (geringe Testschwierigkeit; vgl. Abschnitt 6.2.1 für eine genauere Beschreibung der resultierenden Testversionen). Auch die subjektive Wahrnehmung der Testschwierigkeit wurde nach dem Mathematiktest zum Zwecke der Manipulationskontrolle erfragt. Die Testpersonen schätzten dazu auf einer vierstufigen Likertskala ein, wie schwierig sie den Test fanden.

Aus der Kombination der drei 2-stufig variierten unabhängigen Variablen ergeben sich acht verschiedene Testbedingungen. Die Testpersonen wurden den Testbedingungen randomisiert zugewiesen. Die Randomisierung fand dabei auf Personenebene statt, so dass keine Konfundierung der Versuchsbedingungen mit Schulklassen oder gar Schularten entstehen konnte. In Tabelle 6.3 ist eine Übersicht über das experimentelle Design der Studie mit Angaben zu den jeweiligen Zellenbesetzungen dargestellt[5].

[5] Da es drei verschiedene FIT-Versionen gibt, um Positions- und Inhaltseffekte einzelner Aufgaben einzuschränken (vgl. Abschnitt 6.2.1), resultieren streng genommen 24 verschiedene Testvarianten, denen die Testpersonen randomisiert zugewiesen wurden. Da die drei FIT-Versionen jedoch keinen inhaltlichen Zweck erfüllen und für die Hypothesenprüfungen nicht unterschieden werden, werden diese Testvarianten hier nicht separat aufgeführt.

*Tabelle 6.3*: Experimentelles Design der Studie und Verteilung der Stichprobe auf die Testbedingungen.

| UV1: Testalgorithmus | UV2: Transparenz der Testinstruktion | UV3: Testschwierigkeit | | |
|---|---|---|---|---|
| | | Gering | Mittel | Zeilensumme |
| FIT – CAT | Nein | 83 | 85 | 168 |
| | Ja | 90 | 87 | 177 |
| CAT – FIT | Nein | 92 | 87 | 179 |
| | Ja | 92 | 87 | 179 |
| Spaltensumme | | 357 | 346 | 703 |

*Anmerkungen.* UV: unabhängige Variable; FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

Das Fähigkeitsselbstkonzept geht als Moderatorvariable in die Analysen ein, da angenommen wird, dass es einen Einfluss auf den Zusammenhang zwischen Testalgorithmus und Motivation zur Testbearbeitung ausübt (vgl. Abschnitte 3.2.2.1 und 3.2.2.4). Das dispositionelle Leistungsmotiv wird als Kontrollvariable berücksichtigt. Die abhängige Variable stellt die Motivation zur Testbearbeitung in Form der Erwartungs- und der Wertkomponente dar, wobei vor allem die Erwartungskomponente im Rahmen der untersuchten Fragestellungen von Interesse ist (vgl. Abschnitt 5).

Für die Konzeption des experimentellen Designs waren drei Grundgedanken ausschlaggebend: Erstens sollten Personen- und Testmerkmale berücksichtigt werden. Diesem Anspruch wird genügt, da einerseits das Leistungsmotiv und die motivationalen Überzeugungen als Personenmerkmale berücksichtigt werden und andererseits Testalgorithmus, Transparenz der Testinstruktion und Testschwierigkeit als Testmerkmale variiert werden. Zweitens sollte die vermittelnde Rolle des Fähigkeitsselbstkonzepts genau analysiert werden. Dies geschieht durch die Berücksichtigung des Fähigkeitsselbstkonzepts als Moderator zwischen Testalgorithmus und Motivation zur Testbearbeitung. Drittens sollte die aktuelle Motivation im Handlungsverlauf abgebildet werden. Daher werden mehrere Messzeitpunkte berücksichtigt (vgl. Abschnitt 6.4).

## 6.4 Versuchsdurchführung

Zur Vorbereitung der Studie fand im August 2009 eine Testleiter-Schulung statt. Vier erfahrene studentische Testleiterinnen, die über das DPC in Hamburg rekrutiert worden waren, wurden mit Inhalt und Ziel sowie mit dem Ablauf der Studie im Detail vertraut gemacht. Um eine reibungslose Durchführung der Studie und eine hohe Objektivität zu gewährleisten, erhielten die Testleiterinnen ein Manual mit den wichtigsten organisatorischen Informationen

- zur Testvorbereitung (Materialien, Zeitplan, Richtlinien, Kooperation mit der Schule etc.),
- zur Testdurchführung (Vorbereitungen am Testtag, Aufbau der Materialien im Testraum etc.) und
- zum Verhalten nach der Testung (Ausfüllen des Protokolls, Datensicherung, Abbau der Netbooks etc.).

Außerdem bekamen die Testleiterinnen ein Skript, das einige wenige, bei der Testung mündlich vorzulesende Instruktionen im Wortlaut enthielt, Informationen zum Umgang mit schwierigen Situationen bereitstellte (z. B. bei Unruhe, Zwischenfragen, frühzeitigem Fertigwerden einzelner Jugendlicher) sowie eine detaillierte technische Anleitung zur Bedienung des Computerprogramms beinhaltete (inklusive Angaben zum Verhalten bei typischen Fehlermeldungen). Zur Dokumentation der Testungen wurden den Testleiterinnen Teilnahmelisten und Testleiterprotokolle ausgehändigt, die während beziehungsweise nach den Testungen ausgefüllt wurden. Alle Testleiterinnen unterzeichneten eine Vertraulichkeitserklärung.

*Tabelle 6.4*: Versuchsablauf für die erste und zweite Teststunde.

| Schritt | Inhalt | UV-Manipulation | Dauer in min (ca.) | Dauer kumulativ |
|---|---|---|---|---|
| 1. Teststunde | | | | |
| 1. | Begrüßung, allgemeine Hinweise | | 3 | 3 |
| 2. | Fragebogen Fähigkeitsselbstkonzept, Selbstwirksamkeit | | 4 | 7 |
| 3. | Instruktion Test 1 | intransparent/ transparent | 2 | 9 |
| 4. | Aufgabenbeispiele | | 4 | 13 |
| 5. | Test 1 | FIT/CAT Schwierigkeit gering/mittel | 20 | 33 |
| 6. | Nach 5. Aufgabe: Frage zur wahrgenommenen Leistung, OMQ Teil 1 | | 4 | 37 |
| 7. | Frage zur wahrgenommenen Leistung, OMQ Teil 2, SOS | | 8 | 45 |
| Pause | | | 5 | 50 |
| 2. Teststunde | | | | |
| 8. | Instruktion Test 1 | intransparent/ transparent | 2 | 52 |
| 9. | Test 2 | CAT/FIT Schwierigkeit gering/mittel | 20 | 72 |
| 10. | Nach 5. Aufgabe: Frage zur wahrgenommenen Leistung, OMQ Teil 1 | | 4 | 76 |
| 11. | Frage zur wahrgenommenen Leistung, OMQ Teil 2, SOS | | 8 | 84 |
| 12. | AMS-R | | 4 | 88 |
| 13. | Personenbezogene Angaben | | 2 | 90 |
| 14. | Dank | | 1 | 91 |
| 15. | Ergebnisrückmeldung (optional) | | 4 | 95 |

*Anmerkungen.* OMQ: On-Line Motivation Questionnaire; SOS: Student Opinion Scale; AMS-R: Kurzform der revidierten Achievement Motive Scale; FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

Die Durchführung der Studie fand im September und Oktober 2009 statt, also zwischen den Sommer- und den Herbstferien in Schleswig-Holstein. In der Regel standen für jede Testung zwei Testleiterinnen zur Verfügung. Die Testungen fanden in Klassen- oder Fachräumen der teilnehmenden Schulen statt. Die Netbooks wurden von den Testleiterinnen mitgebracht, vor Beginn der Testung aufgebaut und gestartet. Jede Testung dauerte zwei Schulstunden (2 x 45 min mit einer fünfminütigen Pause zwischen den beiden Stunden; vgl. Übersicht in Tabelle 6.4). In der ersten Teststunde erfolgte eine durch das Testleiterskript vorgegebene kurze mündliche Begrüßung der Schülerinnen und Schüler durch die Testleiterinnen. Alle weiteren Schritte fanden am Netbook statt. Zunächst lasen die Jugendlichen einige allgemeine Hinweise zum Umgang mit dem Computerprogramm und zur Bedienung des Netbooks durch. Das Anklicken der Antwortfelder mit der Maus wurde an einem Beispiel geübt. Die Testpersonen wurden darauf hingewiesen, dass alle Daten vertraulich behandelt werden und die Klassenlehrkraft über die Ergebnisse einzelner Schülerinnen und Schüler nicht informiert würde. Nach diesen allgemeinen Hinweisen gab es Gelegenheit, offene Fragen mit den Testleiterinnen zu klären.

Anschließend begann die eigentliche Testung. Der Start war an das Drücken einer bestimmten Taste geknüpft, was die Jugendlichen vorab nicht wussten. So wurde sichergestellt, dass alle Schülerinnen und Schüler gleichzeitig anfingen. Zunächst wurden das mathematische Fähigkeitsselbstkonzept und die Selbstwirksamkeitserwartung in Mathematik erfasst. Die A-Priori-Messung dieser motivationalen Überzeugungen, mit denen die Jugendlichen in die Testung gehen, stellt sicher, dass diese Einschätzungen nicht durch die folgenden UV-Manipulationen verfälscht sind. Es folgte die eigentliche Testinstruktion. Anschließend wurden zwei Aufgabenbeispiele gegeben, die die Jugendlichen zur Probe beantworten konnten. Diese Ergebnisse gingen nicht in die Kompetenzschätzung ein, sondern dienten lediglich dazu, die Jugendlichen mit dem Aufgabenformat und der Beantwortung vertraut zu machen. Nach den Beispielen begann der Mathematiktest. Nach der Hälfte, also nach fünf Aufgaben, wurde der Test kurz unterbrochen. Es wurde die Frage nach der wahrgenommenen Leistung gestellt und der erste Teil des OMQ vorgegeben, um die aktuelle Motivation im Handlungsverlauf zu erfassen. Anschließend wurde der Mathematiktest durch die sukzessive Vorgabe von weiteren fünf Aufgaben fortgesetzt. Nach dem Test folgten die nochmalige Frage zur wahrgenommenen Leistung, die Bearbeitung des zweiten Teils des OMQ sowie die Vorgabe des SOS, um die aktuelle Motivation unmittelbar nach der Testsituation zu messen. Damit war die erste Teststunde beendet.

Nach einer fünfminütigen Pause begann die zweite Teststunde. Der Ablauf der zweiten Teststunde entsprach weitgehend dem der ersten Teststunde: Es wurde zunächst die Instruktion zum zweiten Mathematiktest gegeben, bevor dieser Test begann. Der Mathematiktest wurde nach fünf Aufgaben unterbrochen, um die wahrgenommene Leistung einzuschätzen und den ersten Teil des OMQ zu bearbeiten. Nach Ende des zweiten Mathematiktests wurden die Frage zur wahrgenommenen Leistung, der zweite OMQ-Teil und der SOS vorgegeben. Außerdem beantworteten die Jugendlichen nun die AMS-R-Kurzform zur Erfassung des Leistungsmotivs sowie die personenbezogenen Fragen. Diese wurden am Ende der Testung gestellt, damit sie nicht zur Verzerrung der Ergebnisse führen können (Braun, Woodley, Richardson & Leidner, 2011). Die Testung schloss ab, indem sich die Jugendlichen auf Wunsch ihre Ergebnisauswertung für beide Teststunden in Form der Anzahl korrekt gelöster Mathematikaufgaben im FIT und im CAT anzeigen lassen konnten. Fast alle Jugendlichen machten von diesem Angebot Gebrauch. Nach Abschluss der Gesamtstudie wurde den jeweiligen Klassenlehrkräften eine Rückmeldung zugeschickt, die die

Ergebnisse der Klasse im Vergleich zu den anderen teilnehmenden Klassen derselben Schulart sowie einen allgemeinen Vergleich der Schularten beinhaltete (vgl. Abschnitt 6.1). Pro Teststunde (45 min) entfielen etwa 15 bis 20 Minuten auf die Bearbeitung des Mathematiktests. In Tabelle 6.4 wird der gesamte Versuchsablauf im Überblick dargestellt.

Der Überblick über die Versuchsdurchführung macht deutlich, dass neben einer A-priori-Auskunft zum Fähigkeitsselbstkonzept und zur Selbstwirksamkeitserwartung mehrere Messzeitpunkte für die aktuelle Motivation zur Testbearbeitung im Handlungsverlauf realisiert wurden. Dies erschien angesichts der theoretisch-konzeptionellen Überlegungen vielversprechend, da es sich bei der aktuellen Motivation zur Testbearbeitung um ein zeitlich instabiles, flüchtiges Merkmal handelt.

## 6.5 Statistische Analyse

Die statistische Analyse der Daten beinhaltet verschiedene Schritte, die in diesem Abschnitt erläutert werden. Zunächst wird beschrieben, wie die Kompetenztestdaten aus den Mathematiktests der ersten und der zweiten Teststunde in Vorbereitung der Hypothesentests der ersten Fragestellung skaliert werden (Abschnitt 6.5.1). Es folgt die Erläuterung der Modellgeltungstests zum Erwartung-Wert-Modell der Motivation zur Testbearbeitung, anhand derer die Hypothesen der ersten Fragestellung untersucht werden (Abschnitt 6.5.2). Anschließend wird dargestellt, wie die Voraussetzungen für die Analyse der Messwiederholungsdaten geprüft werden (Test auf Carryover-Effekte, Prüfung der faktoriellen Invarianz; Abschnitt 6.5.3). Schließlich werden Mehrebenen-Wachstumsmodelle als Methode zur Analyse von Messwiederholungsdaten beschrieben (Abschnitt 6.5.4). Anhand dieser Modelle werden die Hypothesen der zweiten Fragestellung geprüft.

### 6.5.1 Skalierung der Kompetenztests

Um die Verteilung der Personenparameterschätzer zur mathematischen Kompetenz der Schülerinnen und Schüler und die Schwierigkeiten der verwendeten Aufgaben auf einer gemeinsamen Skala abbilden zu können, werden die Rohdaten aus den Kompetenztests skaliert. Der Skalierung wird das eindimensionale Rasch-Modell zugrunde gelegt. Für die vorliegende Stichprobe werden die Aufgabenschwierigkeiten nicht frei geschätzt, sondern mit den Aufgabenschwierigkeiten der Bildungsstandards-Stichprobe verankert (vgl. Abschnitt 6.2.1). Das bedeutet, dass die Aufgabenschwierigkeiten auf die Werte der Referenz-Skalierung fixiert werden. Durch diese so genannte Kalibrierung wird erreicht, dass die Personenparameterverteilungen der Bildungsstandards-Stichprobe und der Stichprobe der vorliegenden Arbeit direkt vergleichbar sind (für nähere Erläuterungen dieser Kalibrierungsmethode siehe de Ayala, 2009).

Die Güte der Skalierung wird anhand des Aufgaben-Fit-Kriteriums *Weighted Mean Square* (WMNSQ; gewichtetes Abweichungsquadrat; vgl. Wright & Linacre, 1994) und dessen *t*-Wertes sowie anhand der EAP/PV-Reliabilität beurteilt. Der WMNSQ einer Aufgabe gibt die Übereinstimmung von

erwarteter und beobachteter Lösungshäufigkeit unter Annahme der Gültigkeit des verwendeten IRT-Modells (hier: des Rasch-Modells) an. Er ist ein Indikator dafür, wie gut eine Aufgabe zwischen Testpersonen unterschiedlicher Kompetenz trennt. Der WMNSQ kann Werte zwischen Null und Unendlich annehmen. Er hat einen Erwartungswert von 1. In diesem Falle passt das postulierte Modell exakt auf die empirischen Daten. Werte größer als 1 indizieren eine schlechte Modellpassung, das heißt, die Daten haben eine geringere Vorhersagekraft als vom Modell erwartet. Ein WMNSQ von 1.4 beispielsweise bedeutet, dass die empirischen Daten 40 Prozent mehr prädiktive Unsicherheit (Zufallseinfluss, „Rauschen“) beinhalten als vom Modell erwartet. Werte kleiner als 1 weisen hingegen auf eine „zu gute“ Modellanpassung hin. Dies bedeutet, dass die Vorhersagekraft der Daten größer ist als vom Modell erwartet. So drückt ein WMNSQ von 0.7 eine Verminderung der vom Modell erwarteten prädiktiven Unsicherheit um 30 Prozent aus. Die Daten verhalten sich sozusagen weniger probabilistisch als erwartet. Zufallsbedingt kann der WMNSQ trotz guter Modellpassung von 1 abweichen. Erst wenn sich der WMNSQ außerhalb des 95-Prozent-Konfidenzintervalls befindet, ist von einer ungenügenden Passung einer Aufgabe zu sprechen. Dies wird durch den zum WMNSQ gehörigen *t*-Wert indiziert, der in diesem Fall einen Betrag größer als 1.96 annimmt.

Die EAP/PV-Reliabilität gibt den Anteil der durch das geschätzte Modell aufgeklärten Varianz an der Gesamtvarianz an und ist ähnlich zu interpretieren wie die interne Konsistenz (Adams, 2005; Rost, 2004). In der Regel ist es daher wünschenswert, dass die EAP/PV-Reliabilität einen Wert von mindestens .70 annimmt. Allerdings ist dieser Wert lediglich als grobe Orientierung zu verstehen, nicht als verbindlicher Cut-off-Wert (z. B. Schmitt, 1996).

Um die Äquivalenz der per Papier und Bleistift erhobenen Bildungsstandards-Daten und der per Computer erhobenen Daten der vorliegenden Studie darüber hinaus abzusichern, wird zusätzlich eine Skalierung der Kompetenztestdaten ohne Verankerung der Aufgabenschwierigkeiten vorgenommen. In dieser Skalierung werden alle Aufgabenschwierigkeiten frei geschätzt, und der Mittelwert der Personenparameterverteilung wird auf Null festgesetzt. Die Güte dieser freien Skalierung wird mittels eines Likelihood-Quotiententests mit der Güte der Skalierung verglichen, in der die Aufgabenschwierigkeiten mit denen aus der Bildungsstandards-Stichprobe verankert sind. Ergibt sich keine signifikante Differenz in der Güte der beiden Skalierungen, ist dies ein weiterer Beleg für die Äquivalenz der Modelle. Die Daten der ersten und der zweiten Teststunde werden separat skaliert und kalibriert, weil es sich um zwei voneinander unabhängige Tests handelt.

### 6.5.2 Modellgeltungstests

Für die Prüfung der Hypothesen zur ersten Fragestellung (Hypothesen 1 bis 4; vgl. Abschnitt 5) wird das in Abschnitt 3.2.2.4 vorgestellte Erwartung-Wert-Modell der Motivation zur Testbearbeitung in ein lineares Strukturgleichungsmodell mit latenten Variablen übersetzt (vgl. Reinecke, 2005). Zunächst wird die globale Modellgeltung überprüft (Hypothese 1). Dies erfolgt unter Orientierung an den von Schermelleh-Engel, Moosbrugger und Müller (2003) angegebenen Cut-off-Kriterien für Fit-Indizes (vgl. auch Bühner, 2006; Hu & Bentler, 1999). Um von einer guten (akzeptablen) Modellanpassung zu sprechen, sollten folgende Kriterien erfüllt sein: $\chi^2/df \leq 2$ ($\leq 3$), RMSEA $\leq .05$ ($\leq .08$), SRMR $\leq .05$ ($\leq .10$), CFI $\geq .97$ ($\geq .95$; für eine detaillierte Beschreibung der Fit-Indizes sei auf Schermelleh-Engel et al., 2003, und Bühner, 2006, verwiesen). Da diese Kriterien lediglich deskriptive

Hinweise auf die Modellgeltung liefern und zu divergierenden Bewertungen führen können, sollte die Beurteilung der Modellgeltung anhand des Gesamteindruckes aller Fit-Indizes vorgenommen werden. Auf die Interpretation des inferenzstatistischen $\chi^2$-Tests, der die Abweichung der modellimplizierten Kovarianzmatrix von der Stichprobenkovarianzmatrix gegen Null testet, wird verzichtet. Der $\chi^2$-Wert ist von der Stichprobengröße abhängig, weshalb der $\chi^2$-Test insbesondere bei großen Stichproben irreführend sein kann. Schermelleh-Engel et al. (2003) empfehlen aufgrund der eingeschränkten Aussagekraft des $\chi^2$-Tests, diesem Test zur Beurteilung der Modellgeltung nicht allzu viel Bedeutung beizumessen und sich stattdessen an den deskriptiven Fit-Indizes zu orientieren.

Anschließend werden einzelne Pfadkoeffizienten genauer analysiert (Hypothese 2) und gegeneinander getestet (Hypothese 3). Schließlich werden zwei alternative, ineinander verschachtelte Modelle (*nested models*) hinsichtlich der Güte ihrer Anpassung an die Daten miteinander verglichen (Hypothese 4). Dieser Modellvergleich erfolgt, in Anlehnung an das von Muthén und Muthén beschriebene Procedere (verfügbar unter http://www.statmodel.com/chidiff.shtml), über einen so genannten $\chi^2$-Differenzentest mit Satorra-Bentler-Korrektur. Mit diesem Test wird die Anpassung eines unrestringierten Modells mit der Anpassung eines restringierten Modells verglichen. Die Satorra-Bentler-Korrektur in Form einer Mittelwertadjustierung der $\chi^2$-Werte wird vorgenommen, um eine gegebenenfalls vorhandene Abweichung der Verteilung der $\chi^2$-Werte von der Normalverteilung auszugleichen (Satorra & Bentler, 1999).

### 6.5.3 Analyse von Messwiederholungsdaten: Prüfung der Voraussetzungen

Um für die Hypothesentests zur zweiten Fragestellung die vollständige Stichprobe ausschöpfen und Mehrebenen-Wachstumsmodelle nutzen zu können (Hypothesen 5 bis 10; vgl. Abschnitt 5), müssen folgende Voraussetzungen zur Analyse von Messwiederholungsdaten erfüllt sein: Zum einen muss ausgeschlossen werden, dass Carryover-Effekte bestehen (Abschnitt 6.5.3.1), zum anderen muss mindestens starke faktorielle Invarianz der verwendeten Variablen vorliegen (Abschnitt 6.5.3.2).

#### *6.5.3.1 Test auf Carryover-Effekte*

Um ausschließen zu können, dass zwischen der ersten und der zweiten Teststunde *Carryover-Effekte* bestehen, werden Multigruppenanalysen durchgeführt. Carryover-Effekte sind unerwünschte Effekte, die in Innersubjektdesigns auftreten können (vgl. Abschnitt 6.3). Sie äußern sich darin, dass die Wirkung der ersten UV-Stufe der Innersubjektvariable (im vorliegenden Fall: des Testalgorithmus) auf die abhängige Variable (AV; im vorliegenden Fall: die Motivation zur Testbearbeitung) die Wirkung der nachfolgenden UV-Stufe auf die AV beeinflusst (Keppel & Wickens, 2004). Der Effekt der ersten UV-Stufe strahlt sozusagen noch auf die zweite UV-Stufe aus und verfälscht deren Wirkung auf die AV. Im vorliegenden Fall könnte dies zum Beispiel bedeuten, dass der Effekt von CAT auf die Motivation zur Testbearbeitung unterschiedlich ausfällt, je nachdem, ob vorab ein FIT bearbeitet wurde oder nicht. Weitere differentielle Effekte je nach Testbedingung sind denkbar. Carryover-Effekte sind nicht mit Reihenfolge-Effekten zu verwechseln, die sich aus der bloßen Abfolge von UV-Stufen ergeben: So ist es plausibel anzunehmen, dass die Motivation in dem ersten Test, der bearbeitet wird, höher ist als in dem zweiten Test, unabhängig davon, ob zuerst ein CAT oder ein FIT bearbeitet wurde. Gründe hierfür können beispielsweise Müdigkeit oder nachlassende Konzentration

sein. Reihenfolge-Effekte wirken gleichermaßen und in inhaltlicher Unabhängigkeit von den UV-Stufen auf die AV ein. Mögliche Verzerrungen in den Daten durch Reihenfolge-Effekte können durch eine Ausbalancierung des Designs, also durch eine über die Testbedingungen hinweg ausgeglichene Abfolge der Stufen der Innersubjektvariable vermieden werden. Dies wurde im Design der vorliegenden Studie berücksichtigt (vgl. Abschnitt 6.3). Auf diese Weise verteilen sich die Reihenfolge-Effekte gleichermaßen über alle Testbedingungen und können die Ergebnisse nicht systematisch verzerren. Carryover-Effekte hingegen hängen inhaltlich von den UV-Stufen ab und können je nach UV-Stufen-Kombination, das heißt, je nach Testbedingung, unterschiedlich ausfallen. Während Reihenfolge-Effekte Verzerrungen im Sinne eines Haupteffekts des Messzeitpunkts auf die Veränderung der AV darstellen, sind Carryover-Effekte im Sinne von Interaktionseffekten zwischen Messzeitpunkt und Innersubjektfaktor auf die AV zu interpretieren. Gerade dies macht Carryover-Effekte, wenn sie bestehen, zu einem ernstzunehmenden Problem für die Interpretation der Daten in Innersubjektdesigns. Liegen Carryover-Effekte in Innersubjektdesigns vor, schlagen Keppel und Wickens (2004) vor, lediglich die Daten des ersten Messzeitpunkts für die Analysen zu nutzen.

Die Daten der vorliegenden Arbeit werden mit Hilfe von Multigruppenanalysen auf Carryover-Effekte überprüft: Die Veränderung der AV von der ersten zur zweiten Teststunde wird zwischen den beiden Gruppen, die sich in der Reihenfolge der Bearbeitung von CAT und FIT unterscheiden (d. h. dadurch, ob zuerst der CAT und dann der FIT oder zuerst der FIT und dann der CAT bearbeitet wurde), verglichen und auf statistische Signifikanz geprüft. Zeigt sich innerhalb der beiden Gruppen eine signifikante Veränderung der AV über die Zeit, liegt ein Reihenfolge-Effekt vor, der aufgrund des ausbalancierten Designs für die Beantwortung der Fragestellungen dieser Arbeit irrelevant ist. Fällt der Reihenfolge-Effekt jedoch zwischen den beiden Gruppen unterschiedlich aus, spricht dies für einen Carryover-Effekt, also für eine systematisch verzerrte Veränderung der AV in Abhängigkeit davon, welche UV-Stufe zuerst angewendet wurde.

#### *6.5.3.2 Prüfung der faktoriellen Invarianz*

Für eine sinnvolle Analyse von Messwiederholungsdaten sollte darüber hinaus mindestens starke Messinvarianz vorliegen, was bedeutet, dass die Faktorenstruktur, die Faktorladungen und die Intercepts über die Zeit hinweg konstant bleiben (Geiser, 2010). Vor Beginn der eigentlichen Modellierung wird daher überprüft, ob die faktorielle Invarianz der Motivationsdaten aus den beiden Messzeitpunkten bestätigt werden kann. Nur wenn sich die genannten Parameter zwischen den beiden Messzeitpunkten nicht signifikant voneinander unterscheiden, ist davon auszugehen, dass sich die psychometrischen Eigenschaften der wiederholt verwendeten Messinstrumente nicht verändert haben und die latenten Merkmale der beiden Teststunden hinreichend gut miteinander vergleichbar sind. Die Prüfung der faktoriellen Invarianz erfolgt über $\chi^2$-Differenzentests, in welchen sukzessive die Anpassung von Modellen mit zunehmend stärkeren Gleichheitsrestriktionen mit der Anpassung des allgemeinen unrestringierten Modells verglichen wird.

### 6.5.4 Analyse von Messwiederholungsdaten: Mehrebenen-Wachstumsmodelle

Nachdem die Daten auf das Vorhandensein von Carryover-Effekten und auf faktorielle Invarianz hin überprüft worden sind, finden die eigentlichen Hypothesentests zur zweiten Fragestellung statt. Um das experimentelle Innersubjekt-Design angemessen statistisch abbilden zu können, werden die

Hypothesen 5 bis 10 mittels Mehrebenen-Wachstumsmodellen getestet (vgl. Luke, 2004; Muthén & Muthén, 2009). Diese Modelle eignen sich dafür, Messwiederholungsdaten komplexer Stichproben mit einem regressionsanalytischen Ansatz zu untersuchen (Little, Schnabel & Baumert, 2000; Skrondal & Rabe-Hesketh, 2004). Zunächst ist es wegen des counterbalancierten Designs erforderlich, die Datenmatrix nach dem so genannten Wide-To-Long-Ansatz zu strukturieren, so dass die Daten der beiden Messzeitpunkte pro Testperson in zwei Zeilen erscheinen (Muthén & Asparouhov, 2009; Muthén & Muthén, 2009). Ist mindestens starke faktorielle Invarianz zwischen den Messzeitpunkten gegeben, wird das Mehrebenen-Wachstumsmodell so spezifiziert, dass auf der untersten Ebene die Zeit modelliert wird, also die beiden Messzeitpunkte der ersten und der zweiten Teststunde (Ebene 1). Auf Ebene 2 werden die einzelnen Schülerinnen und Schüler modelliert (vgl. Muthén & Asparouhov, 2009). Auf diese Weise werden die Abhängigkeiten der geschachtelten Daten, das heißt, die zwei Messzeitpunkte pro Person und mögliche Reihenfolgeeffekte, angemessen berücksichtigt. Zudem können die UVs auf den jeweiligen Ebenen modelliert werden, auf denen sie variiert wurden. Während der Testalgorithmus als Innersubjektfaktor auf Ebene 1 modelliert wird, gehen die Testinstruktion und die Testschwierigkeit als Ebene-2-Variablen in das Modell ein. Über die Definition von festen und zufälligen Effekten können Interaktionseffekte wahlweise innerhalb einer Ebene oder als so genannte Zwischen-Ebenen-Interaktionen modelliert und getestet werden (vgl. Abschnitte 7.3 bis 7.6).

#### *6.5.4.1 Modellierung der Stichprobenstruktur*

Die komplexe Stichprobenstruktur, die sich aus der Tatsache ergibt, dass Jugendliche in Klassen in Schularten getestet wurden, wird in den Mehrebenen-Wachstumsmodellen über entsprechende Stratifizierungs- und Clusterdefinitionen modelliert. Als Stratifizierungsvariable wird die Schulart verwendet, da die Stichprobenziehung, unter Orientierung an der Populationsverteilung auf die Schularten, innerhalb der Schularten erfolgt war (vgl. Abschnitt 6.1). Als Clustervariable dient die Klassenzugehörigkeit der Jugendlichen. Die Modellierung der komplexen Stichprobenstruktur trägt bei der Modellschätzung der Tatsache Rechnung, dass sich Elemente innerhalb einer Einheit, beispielsweise Jugendliche innerhalb einer Klasse, stärker ähneln als Elemente verschiedener Einheiten oder als Elemente, die völlig unabhängig voneinander sind. Eine Nichtbeachtung dieser Abhängigkeiten in den Daten kann wegen einer Überschätzung der effektiven Stichprobengröße zu einer Unterschätzung der Standardfehler der Modellparameter führen und so progressiv verzerrte Signifikanztests bewirken (für nähere Ausführungen hierzu siehe Geiser, 2010).

#### *6.5.4.2 Beurteilung der Modellgüte*

Um die Angemessenheit und die Güte der Mehrebenen-Wachstumsmodelle beurteilen zu können, werden pro Modell der Intraklassen-Koeffizient (ICC) und ein Maß für die Varianzaufklärung ($R^2$) angegeben. Der ICC gibt im vorliegenden Fall an, wie hoch der Anteil der Varianz in der abhängigen Variable ist, der sich auf die Elemente der Ebene 2 zurückführen lässt (vgl. Luke, 2004). Die Bestimmung des ICC erfolgt anhand des Nullmodells, das heißt, anhand eines Modells, das keine Prädiktoren enthält. Der ICC ist wichtig, um die Angemessenheit der Mehrebenen-Modellierung zu prüfen. So wäre es unpassend, eine Mehrebenenstruktur zu modellieren, wenn der ICC Null beträgt. Denn dies würde bedeuten, dass durch die Elemente auf Ebene 2 keine Varianz in der abhängigen Variable erklärt wird. In der vorliegenden Arbeit gibt der ICC an, welcher Varianzanteil in der Motivation zur Testbearbeitung auf Unterschiede zurückgeht, die in den Personen begründet liegen (Ebene 2). Der Komplementäranteil des ICC zu 1 (d. h. 1-ICC) gibt an, wie groß die Varianz in der

Motivation zur Testbearbeitung anteilig ist, die auf den Messzeitpunkt zurückgeht (Ebene 1). Da sich in den hier berichteten Mehrebenen-Wachstumsmodellen die Personen auf Ebene 2 befinden (statt wie in Querschnittsstudien auf Ebene 1), sind recht hohe ICC-Werte zu erwarten.

Die Beurteilung der Modellgüte gestaltet sich in Mehrebenenmodellen mit zufälligen Effekten im Vergleich zu „einfachen" Regressionsmodellen ohne Mehrebenenstruktur schwierig (Luke, 2004). In einfachen Regressionsmodellen gibt $R^2$ den Anteil an der Varianz in der abhängigen Variable an, der sich auf die Prädiktoren im Modell zurückführen lässt. In Mehrebenenmodellen gibt es ein $R^2$ pro Ebene, und die herkömmliche Berechnung von $R^2$ kann bei dem Vorhandensein zufälliger Effekte zu nicht sinnvoll interpretierbaren Werten führen. Hox (2000) und Luke (2004) schlagen daher unter Bezugnahme auf Snijders und Bosker (1999) für Mehrebenenmodelle mit zufälligen Effekten ein alternatives Vorgehen für die Berechnung von $R^2$ vor, dem in der vorliegenden Arbeit gefolgt wird. Dabei wird $R^2$ als proportionaler Anteil verstanden, um den die Residualvarianz des Nullmodells im vollständigen Modell verringert wird. Allerdings ist zur Herstellung der Vergleichbarkeit der Residualvarianzterme auf Ebene 1 und Ebene 2 zwischen Nullmodell und vollständigem Modell eine Restriktion erforderlich: So müssen die so genannten *random slopes* der hier verwendeten Mehrebenen-Wachstumsmodelle, also die als zufällige Effekte auf Ebene 2 modellierten Steigungen aus Ebene 1, zur Berechnung von $R^2$ als feste Effekte modelliert werden. Andernfalls bestünde die Residualvarianz auf Ebene 2 im vollständigen, unrestringierten Modell nicht nur aus einem Intercept-abhängigen Term (wie im Nullmodell), sondern zusätzlich auch aus prädiktorspezifischen Termen, die jedoch im Nullmodell nicht vorhanden sind. Die Vergleichbarkeit der Residualvarianzen wäre dadurch eingeschränkt. Um zumindest eine Approximation eines Wertes zur Modellgüte zu erhalten, wird daher (nur für die Berechnung von $R^2$) der Kompromiss eingegangen, die eigentlich zufälligen Steigungen in einem restringierten Vergleichsmodell zu fixieren. Es ist davon auszugehen, dass die tatsächliche Modellgüte der für die statistischen Analysen verwendeten unrestringierten Modelle höher liegt als die berichteten $R^2$-Werte, die sich auf das restringierte Modell beziehen. Denn durch die Modellierung der Personen auf Ebene 2 besteht auf dieser Ebene voraussichtlich eine hohe Varianz (sichtbar an hohen ICCs), die zur Berechnung der Modellgüte künstlich verringert wird.

#### *6.5.4.3 Simple-Slope-Analysen*

Beinhaltet ein Mehrebenen-Wachstumsmodell einen signifikanten Interaktionseffekt, wird dieser in einem weiteren Mehrebenen-Wachstumsmodell über Simple-Slope-Analysen inhaltlich untersucht. Hierzu werden für den metrischen Prädiktor des Interaktionsterms individuelle Abweichungswerte vom Mittelwert zuzüglich einer Standardabweichung beziehungsweise vom Mittelwert abzüglich einer Standardabweichung gebildet. Der Effekt des kategorialen Prädiktors auf die Motivation zur Testbearbeitung wird dann für diese beiden Ausprägungen getrennt betrachtet (für Details zu diesem Vorgehen siehe Aiken & West, 1991, sowie Richter, 2007).

### 6.5.5 Verwendete Software und Signifikanzniveau

Die Skalierung der Kompetenztestdaten wird mit der Item-Response-Modellierungs-Software ConQuest 2.0 vorgenommen (Wu, Adams, Wilson & Haldane, 2007), die auch zur ursprünglichen Skalierung der Bildungsstandards-Daten verwendet wurde (Prenzel & Blum, 2007). Alle anderen Analysen und sämtliche Hypothesentests (Modellgeltungstests, Prüfung der Voraussetzungen für die

Analyse von Messwiederholungsdaten, Mehrebenen-Wachstumsmodelle) werden mit der statistischen Modellierungssoftware M*plus*, Version 5 (Muthén & Muthén, 2009) durchgeführt. Allen statistischen Signifikanztests wird eine Irrtumswahrscheinlichkeit von $\alpha = .05$ zugrunde gelegt.

# 7 Ergebnisse

In diesem Abschnitt werden die Ergebnisse der vorliegenden Studie berichtet. Zunächst werden die Resultate der Skalierung und Kalibrierung der mathematischen Kompetenztests zusammengefasst (Abschnitt 7.1). Es folgt die Ergebnisdarstellung zu den Hypothesen: In Abschnitt 7.2 wird das theoretisch postulierte Erwartung-Wert-Modell einer empirischen Prüfung unterzogen, um die Hypothesen 1 bis 4 zum Gesamtmodell sowie zu einzelnen Pfaden des Modells zu untersuchen. In Abschnitt 7.3 werden Ergebnisse zur Eignung der Daten für Messwiederholungsanalysen berichtet. In Abschnitt 7.4 werden die Effekte von Testalgorithmus und Fähigkeitsselbstkonzept auf die Motivation zur Testbearbeitung analysiert und damit die Ergebnisse zu den Hypothesen 5 und 6 präsentiert. Wie sich die Testinstruktion auf die Motivation zur Testbearbeitung auswirkt (Hypothesen 7 und 8), ist Gegenstand von Abschnitt 7.5. Die Ergebnisse zu den Hypothesen 9 und 10, die sich auf den Einfluss der Testschwierigkeit auf die Motivation zur Testbearbeitung beziehen, werden in Abschnitt 7.6 dargestellt.

## 7.1 Skalierung der Kompetenztests

Die Mathematikaufgaben, die in den Kompetenztests verwendet werden, entstammen einem bestehenden Aufgabenpool. Dieser Aufgabenpool diente der Normierung von Testaufgaben zur Messung der Anforderungen, die seitens der Bildungsstandards in Mathematik für den Mittleren Schulabschluss formuliert werden (vgl. Abschnitt 6.2.1). Um die Personenparameter und die Aufgabenschwierigkeiten auf einer gemeinsamen Skala abtragen zu können, wurden die Kompetenztestdaten unter Verwendung des Rasch-Modells für dichotome Daten skaliert (vgl. Abschnitt 2.1). Um außerdem die Personenparameter der vorliegenden Studie mit denjenigen der Bildungsstandards-Stichprobe vergleichen zu können, wurden die Daten an der Bildungsstandards-Stichprobe kalibriert (vgl. Abschnitte 6.2.1 und 6.5). Die Kalibrierung computerbasierter Daten an papierbasierten Daten gilt angesichts der hohen Ökonomie dieses Vorgehens und der in der Regel geringen Effekte des Administrationsmodus auf die Testdaten als weithin akzeptiert (de Ayala, 2009; Lunz et al., 1994; Segall, 2005). Dennoch sind einige Autoren der Auffassung, dass die Äquivalenz der anhand von Papier- und von Computertests gewonnenen Daten gezeigt werden muss (z. B. Green, Bock, Humphreys, Linn & Reckase, 1984). Im vorliegenden Fall kann die Äquivalenz als gegeben angesehen werden, wenn es gelingt, die computerisiert erfassten Kompetenztestdaten der vorliegenden Studie angemessen mit der Bildungsstandards-Stichprobe zu verankern. Dies wird im Folgenden geprüft.

Von den 110 Aufgaben, die für die mathematischen Kompetenztests zur Verfügung standen (vgl. Abschnitt 6.2.1), wurden 89 Aufgaben tatsächlich von mindestens einer Testperson bearbeitet (81 %). Berücksichtigt man, dass jede Testperson 10 der 110 Aufgaben (9 %) im Rahmen des FIT bearbeitet hat, wurden etwa 79 Prozent der verfügbaren Aufgaben im Rahmen des CAT verwendet. Dass etwa drei Viertel der möglichen Aufgaben genutzt wurden, spricht dafür, dass die angestrebte breite Kompetenzverteilung in der Stichprobe realisiert werden konnte. Dennoch wurde nicht jede Aufgabe ausgewählt, was signalisiert, dass mit dem Aufgabenpool sogar noch heterogenere Stichproben untersucht werden könnten. Sieben der 89 verwendeten Aufgaben zeigen keine Varianz

und können daher nicht für die Skalierung verwendet werden. Das Fehlen der Varianz lässt sich darauf zurückführen, dass diese Aufgaben entweder leidglich einer einzigen Testperson vorgegeben wurden oder dass alle Testpersonen dieselbe Antwort gegeben haben. Es verbleiben demnach 82 skalierbare Aufgaben. Da die Skalierung für die beiden Teststunden separat durchgeführt wird, gehen jedoch nicht alle 82 Aufgaben zugleich in die Skalierung ein. Auf Basis der Daten der ersten Teststunde werden 80 Aufgaben skaliert, auf Basis der Daten der zweiten Teststunde 79 Aufgaben.

In einem ersten Schritt werden die Schwierigkeiten der Aufgaben der jeweiligen Teststunde mit den entsprechenden Aufgabenschwierigkeiten der Kalibrierungsstichprobe (Bildungsstandards) verankert, also auf die Werte aus der Kalibrierung fixiert. Die Personenparameter werden in Form von EAPs frei geschätzt. In einem zweiten Schritt wird die Güte des Modells anhand der Aufgaben-Fit-Kriterien WMNSQ und dessen *t*-Wertes überprüft (vgl. Abschnitt 6.5). Neben dem *t*-Wert werden die EAP/PV-Reliabilität und die Varianz der Personenparameter betrachtet.

Bei Fixierung aller Aufgabenschwierigkeiten ergeben sich für die Daten der ersten Teststunde nach 14 Iterationen (Abbruchkriterium: Veränderung der Devianz zwischen den Iterationen < 0.0001) für sechs der 80 Aufgaben (8 %) inakzeptable Anpassungswerte ($t > 1.96$). Daher wird die Schwierigkeit der Aufgabe mit der schlechtesten Anpassung frei gesetzt, und die Daten werden neu skaliert. Es zeigt sich, dass die Schwierigkeit der frei gesetzten Aufgabe in der vorliegenden Studie deutlich geringer ist als in der Kalibrierungsstichprobe. Diese Aufgabe ist für die hier verwendete Stichprobe also einfacher als für die Bildungsstandards-Stichprobe (Anteil korrekter Lösungen: 59 % in der vorliegenden Stichprobe gegenüber 35 % in der Bildungsstandards-Stichprobe). Nach Freisetzung dieser Aufgabenschwierigkeit ergibt sich eine zufriedenstellende Modellanpassung: Lediglich drei Aufgaben (4 %) zeigen weiterhin ungenügende Fit-Werte, was jedoch unter fünf Prozent und damit im tolerablen Zufallsbereich liegt. Für die übrigen 77 Aufgaben ergeben sich gute bis sehr gute Anpassungswerte. Die EAP/PV-Reliabilität liegt mit .67 nur knapp unter dem angestrebten Wert von .70 und ist für Forschungszwecke als ausreichend hoch zu bewerten.

Für die zweite Teststunde führt die Skalierung nach 12 Iterationen (Abbruchkriterium: Veränderung der Devianz zwischen den Iterationen < 0.0001) auf Anhieb zu einem zufriedenstellenden Ergebnis: So zeigen lediglich drei der 79 Aufgaben *t*-Werte größer als 1.96 (4 %), was im Zufallsbereich liegt und somit tolerabel ist. Daher verbleibt es bei der Fixierung aller Aufgabenschwierigkeiten. Die EAP/PV-Reliabilität der Skalierung der zweiten Teststunde liegt ebenfalls bei .67 und damit wie in der Skalierung der Daten der ersten Teststunde nur geringfügig unter dem Richtwert von .70.

Um die Vergleichbarkeit der vorliegenden, computerisiert erhobenen Daten mit den per Papier und Stift erhobenen Daten der Bildungsstandards-Stichprobe inferenzstatistisch abzusichern, wird abschließend ein Likelihood-Quotiententest durchgeführt. Dieser vergleicht die Güte der Skalierung mit den verankerten Aufgabenschwierigkeiten mit der Güte einer freien Skalierung (ohne Verankerung). Es wird also getestet, ob die Modellanpassung durch die Restriktion der Verankerung statistisch bedeutsam schlechter ausfällt als wenn die Parameter des Modells anhand der Daten der vorliegenden Stichprobe frei geschätzt werden. Erfreulicherweise unterscheiden sich die beiden Skalierungen weder bei den Daten aus der ersten Teststunde noch bei den Daten aus der zweiten Teststunde (1. Teststunde: $\Delta\chi^2 = 276.51$, $\Delta df = 78$, $p > .99$; 2. Teststunde: $\Delta\chi^2 = 300.74$, $\Delta df = 78$,

$p > .99$). Diese Ergebnisse unterstützen die Äquivalenz der Daten aus der Bildungsstandards-Stichprobe und der vorliegenden Stichprobe nachhaltig.

Die Skalierungen mit Verankerung der Aufgabenschwierigkeiten an den entsprechenden Parametern aus der Kalibrierungsstichprobe können folglich für die weiteren Analysen verwendet werden. Abbildung 7.1 und Abbildung 7.2 zeigen so genannte Wright Maps zu den Skalierungen mit Verankerung, in denen die geschätzten Personenparameter (Kreuze links der vertikalen Achse) den Aufgabenschwierigkeiten (Nummern rechts der vertikalen Achse) gegenübergestellt werden. Es fällt auf, dass in den Randbereichen der gemeinsamen Skala zwar noch Personen auftreten, denen jedoch kaum Aufgaben gegenüberstehen. Angesichts der Tatsache, dass auch für diese Skalenbereiche Aufgaben im Aufgabenpool zur Verfügung standen, die jedoch nicht oder nur jeweils einer einzelnen Person vorgegeben wurden und daher nicht in die Skalierung eingingen, ist zu vermuten, dass der CAT mit zehn Aufgaben etwas zu kurz war. Die Begrenzung der Testlänge mag dazu geführt haben, dass kaum Aufgaben aus den extremen Schwierigkeitsbereichen unter -2 Logits und über 2 Logits vorgegeben wurden. Ein längerer Test hätte vermutlich zu präziseren Personenparameterschätzungen in den Randbereichen der Verteilung geführt (vgl. auch die Testinformationskurven im Anhang: Abbildung 11.1 und Abbildung 11.2), wäre jedoch im Rahmen der vorliegenden Studie organisatorisch nicht zu realisieren gewesen. Das Hauptinteresse der vorliegenden Arbeit gilt überdies der differenzierten Erfassung der Motivation zur Testbearbeitung und nicht der möglichst genauen Messung der mathematischen Kompetenz.

Bei Betrachtung der Personenparameterverteilung ist auf einen unerwarteten Befund hinzuweisen. So wurde der Mittelwert der Personenparameterverteilung bei der Skalierung der Bildungsstandards-Daten auf Null fixiert. Die Personenparameter in der vorliegenden Studie hingegen wurden frei geschätzt, da die Verankerung über die Aufgabenschwierigkeiten erfolgte. Bei einer ähnlichen Kompetenzverteilung in beiden Stichproben wäre zu erwarten, dass auch der Mittelwert der Personenparameterverteilung der hier verwendeten Stichprobe etwa bei Null liegt. Dies ist nicht der Fall: In der vorliegenden Studie ergibt sich für die Personenparameterverteilung ein Mittelwert von -0.31 (1. Teststunde) beziehungsweise von -0.34 (2. Teststunde). Das bedeutet, dass die mathematische Kompetenz in der Stichprobe der vorliegenden Studie im Durchschnitt etwas niedriger ausfällt als in der Kalibrierungsstichprobe. Dieser Umstand hat Folgen für die Manifestation der Stufen der unabhängigen Variable „Schwierigkeit“ und beeinträchtigt die Prüfung der Hypothesen 9 und 10 (siehe Abschnitt 7.6).

Es ist zusammenzufassen, dass die Skalierung mit Verankerung an der Bildungsstandards-Stichprobe für die vorliegenden Zwecke in zufriedenstellender Weise vorgenommen werden konnte und von einer hinreichenden Äquivalenz der per Papier und Stift beziehungsweise per Computer erhobenen Daten auszugehen ist. Die EAPs aus dieser Skalierung können somit im Folgenden als Personenparameter verwendet werden. Allerdings zeigt sich eine leichte negative Verschiebung der Personenparameterverteilung der vorliegenden Stichprobe gegenüber der Verteilung in der Kalibrierungsstichprobe, was eine Veränderung der beiden Stufen der UV „Schwierigkeit“ nach sich zieht. Dies muss in Abschnitt 7.6 bei der Darstellung der Ergebnisse zu den Hypothesen 9 und 10 berücksichtigt werden.

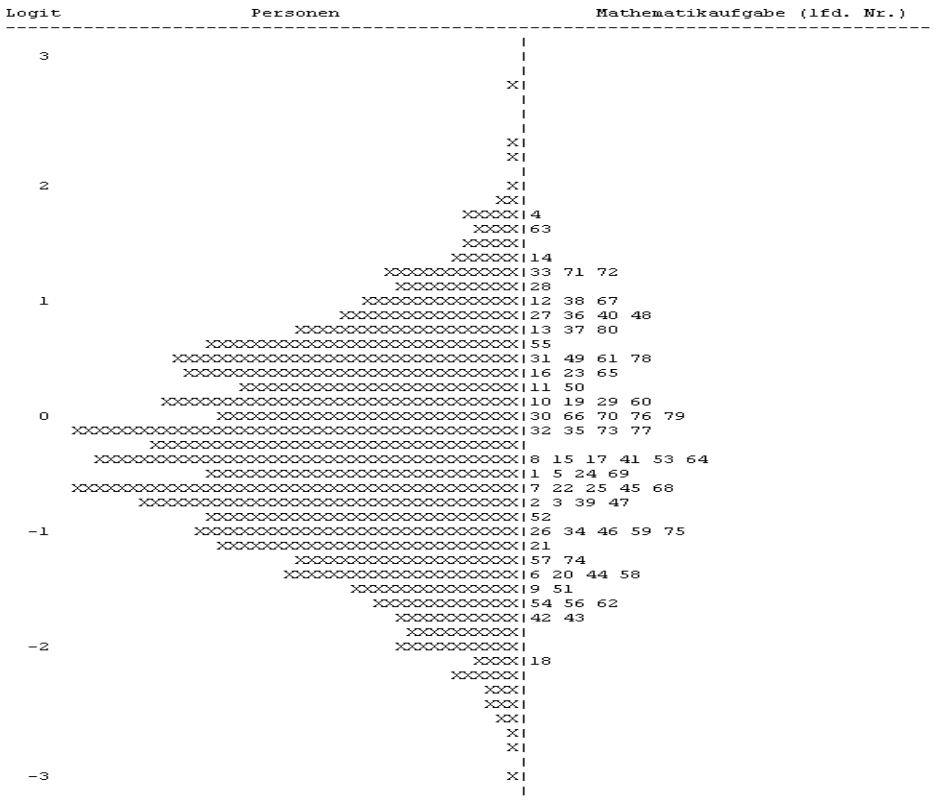



*Abbildung 7.1:* Wright Map zur Gegenüberstellung der Verteilungen der Personenparameter und der Aufgabenschwierigkeiten für die erste Teststunde (ein Kreuz entspricht einer Person; Quelle: ConQuest 2.0).

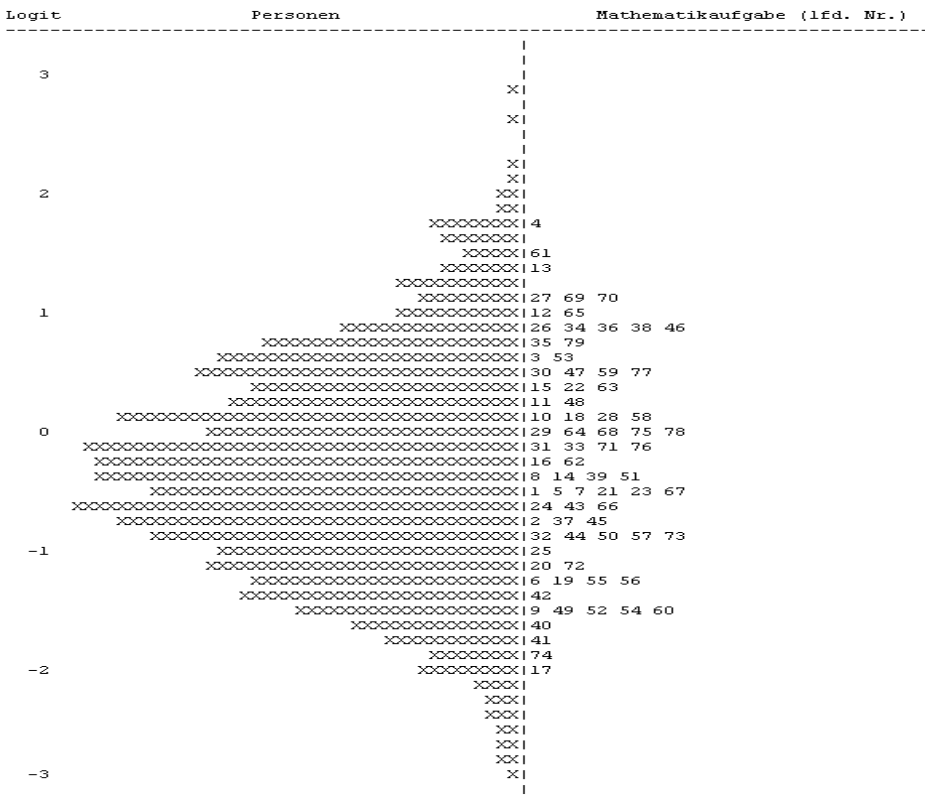



*Abbildung 7.2:* Wright Map zur Gegenüberstellung der Verteilungen der Personenparameter und der Aufgabenschwierigkeiten für die zweite Teststunde (ein Kreuz entspricht einer Person; Quelle: ConQuest 2.0).

## 7.2 Erwartung-Wert-Modell der Motivation zur Testbearbeitung

In diesem Abschnitt werden die Ergebnisse zu der ersten Fragestellung (Hypothese 1 bis 4) dargestellt, die die Eignung des theoretisch abgeleiteten Erwartung-Wert-Modells der Motivation zur Testbearbeitung zum Gegenstand hat. Das Ziel ist, den Prozess der Motivation zur Testbearbeitung angemessen in einem Modell abzubilden. Die Hypothesen werden über lineare Strukturgleichungsmodelle geprüft (vgl. Abschnitt 6.5). Sofern es inhaltlich nützlich erscheint, werden der Präsentation der Befunde eines Hypothesentests relevante deskriptive Ergebnisse vorangestellt. Für eine vollständige Übersicht über die Mittelwerte und Standardabweichungen der per Fragebogen erhobenen Skalen sei auf Tabelle 6.2 im Methodenabschnitt verwiesen.

In Hypothese 1 wird behauptet, dass das theoretisch abgeleitete Erwartung-Wert-Modell der Motivation zur Testbearbeitung (Abbildung 3.4) durch die empirischen Daten insgesamt bestätigt werden kann. Um diese Hypothese zu testen, wird das theoretische Modell in ein empirisches Modell übersetzt. Die Modellgeltung wird für jeden der vier Messzeitpunkte (während und nach Test 1, während und nach Test 2) geprüft. Das Modell muss, wie in Abschnitt 3.2.2.4 dargelegt, bei Bezug auf eine Testsituation der Besonderheit Rechnung tragen, dass die motivationalen Überzeugungen und Dispositionen aus früheren leistungsbezogenen Erfahrungen in die Testsituation mitgebracht werden und somit vorliegen, bevor der aktuelle Prozess der Testbearbeitung beginnt. Daher wird der Einfluss dieser motivationalen Überzeugungen und Dispositionen auf die aktuelle Motivation, die aus der Erwartungs- und der Wertkomponente besteht, direkt modelliert. Als Indikatoren für die motivationalen Überzeugungen werden das Fähigkeitsselbstkonzept und die Selbstwirksamkeitserwartung verwendet, da diese beiden Merkmale diesbezüglich als theoretisch und empirisch vielfach bewährt gelten können. Multikollinearität liegt zwischen den beiden Merkmalen nicht vor (der *Variance Inflation Factor* liegt für alle vier Messzeitpunkte mit Werten von 1.19 bis 1.21 deutlich unter dem kritischen Wert von 10; Eid, Gollwitzer & Schmitt, 2010). Die wahrgenommene Testleistung wirkt sich ebenfalls direkt auf die aktuelle Motivation zur Testbearbeitung aus, die wiederum die aktuelle Testleistung beeinflusst. Die optionalen Pfade der subjektiven Kompetenzüberzeugungen auf die Wertkomponente (vgl. theoretische Erläuterungen in Abschnitt 3.2.2.4) werden im Sinne eines sparsamen Modells nicht berücksichtigt. In Abbildung 7.3 ist das Strukturmodell exemplarisch für den ersten Messzeitpunkt, also während des Tests in der ersten Teststunde, dargestellt. Die latenten Variablen sind, wie bei solchen Modellen üblich, als Ellipsen dargestellt, die manifesten Variablen als Rechtecke. Das dazugehörige Messmodell, das die Faktorladungen der einzelnen Fragebogenitems beinhaltet, sowie die Struktur- und Messmodelle der anderen drei Messzeitpunkte können Abbildung 11.3 bis Abbildung 11.6 im Anhang entnommen werden.

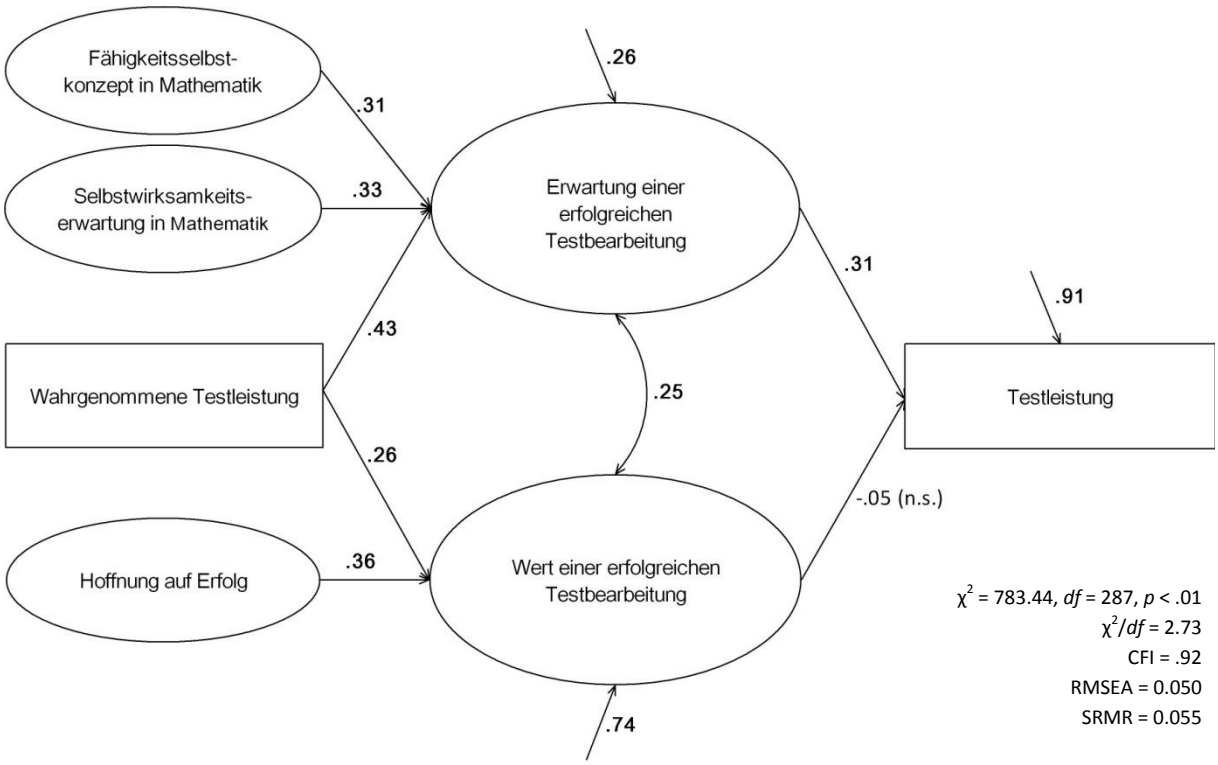



*Abbildung 7.3:* Strukturmodell zur Motivation zur Testbearbeitung während der ersten Teststunde (vollstandardisierte Lösung; signifikante Koeffizienten sind fett gedruckt).

Die in den Strukturgleichungsmodellen verwendete Nomenklatur ist in Tabelle 7.1 dargestellt. Das Superskript „s“ kennzeichnet im Folgenden standardisierte Koeffizienten.

*Tabelle 7.1*: Nomenklatur der Strukturgleichungsmodelle.

| Abkürzung | Bedeutung |
|---|---|
| $\eta$ | Latente endogene Variable |
| $\eta_1$ | Erwartung einer erfolgreichen Testbearbeitung |
| $\eta_2$ | Wert einer erfolgreichen Testbearbeitung |
| $\xi$ | Latente exogene Variable |
| $\xi_1$ | Fähigkeitsselbstkonzept in Mathematik |
| $\xi_2$ | Selbstwirksamkeitserwartung in Mathematik |
| $\xi_3$ | Hoffnung auf Erfolg |
| Y | Manifeste endogene Variable: Testleistung |
| X | Manifeste exogene Variable: Wahrgenommene Testleistung |
| $\zeta$ | Residualvarianz einer latenten endogenen Variable |
| $\zeta_1$ | Residualvarianz der Erwartung einer erfolgreichen Testbearbeitung |
| $\zeta_2$ | Residualvarianz des Werts einer erfolgreichen Testbearbeitung |
| $\varepsilon$ | Residualvarianz der manifesten endogenen Variable Testleistung |
| $\gamma$ | Pfadkoeffizient auf eine latente endogene Variable |
| $\beta$ | Pfadkoeffizient auf eine manifeste endogene Variable |
| $\psi$ | Korrelationskoeffizient zwischen zwei latenten endogenen Variablen |

Bevor zur Prüfung von Hypothese 1 der globale Modellfit der Strukturgleichungsmodelle analysiert wird, wird eine deskriptive Betrachtung der Varianzaufklärung der endogenen Variablen Erwartung, Wert und Testleistung vorgenommen. Durch die Variablen im Modell wird je nach Messzeitpunkt 60 bis 74 Prozent der Varianz der Erwartungskomponente erklärt, während der Anteil erklärter Varianz der Wertkomponente 21 bis 34 Prozent beträgt (Tabelle 7.2). Im Verlauf der Zeit, also von $t_1$ bis $t_4$, nähern sich die Anteile erklärter Varianz etwas an. In Bezug auf die Testleistung ist der durch die aktuelle Motivation erklärte Varianzanteil mit acht bis zwölf Prozent zwar relativ gering, jedoch keineswegs unbedeutend. Dieser Prozentanteil liegt in der theoretisch zu erwartenden und in anderen empirischen Studien gefundenen Höhe (vgl. Abschnitt 4).

*Tabelle 7.2*: Erklärte Varianz der endogenen Variablen des Strukturgleichungsmodells zu den vier Messzeitpunkten.

| | | Anteil erklärter Varianz | | | |
|---|---|---|---|---|---|
| Endogene Variable | Erklärte Varianz | $t_1$ | $t_2$ | $t_3$ | $t_4$ |
| Erwartung einer erfolgreichen Testbearbeitung | $1-\zeta_1$ | .74 | .70 | .68 | .60 |
| Wert einer erfolgreichen Testbearbeitung | $1-\zeta_2$ | .26 | .21 | .27 | .34 |
| Testleistung | $1-\varepsilon$ | .09 | .08 | .12 | .11 |

*Anmerkungen.* $t_1$: während des Tests (1. Teststunde), $t_2$: nach dem Test (1. Teststunde), $t_3$: während des Tests (2. Teststunde), $t_4$: während des Tests (2. Teststunde).

Die Fit-Indizes zur globalen Modellgeltung aller vier Messzeitpunkte (vgl. Hypothese 1) sind in Tabelle 7.3 aufgeführt. Der inferenzstatistische $\chi^2$-Modellgeltungstest weist zwar auf eine signifikante Abweichung der empirischen Kovarianzstruktur von der theoretisch postulierten Kovarianzstruktur hin, doch erscheint dieses Ergebnis aus den in Abschnitt 6.5.2 beschriebenen Gründen wenig aussagekräftig (vgl. Schermelleh-Engel et al., 2003). Die Beurteilung der Modellgeltung erfolgt stattdessen anhand der deskriptiven Fit-Indizes. Der RMSEA bescheinigt zu drei Messzeitpunkten eine gute und zu einem Messzeitpunkt eine akzeptable Anpassung des Modells. Der SRMR und der Quotient $\chi^2/df$ zeigen durchgehend akzeptable Werte. Lediglich der CFI liegt zu allen vier Messzeitpunkten knapp unterhalb des angestrebten Wertes von .95. Eine zusammenfassende Betrachtung der Modellfit-Kriterien weist darauf hin, dass das theoretisch postulierte Erwartung-Wert-Modell der Motivation zur Testbearbeitung durch die empirischen Daten in zufriedenstellender Weise abgebildet wird. Hypothese 1 gilt als bestätigt.

*Tabelle 7.3*: Globale Modellgeltung der linearen Strukturgleichungsmodelle für die vier Messzeitpunkte.

| | | | Fit-Indizes | | | |
|---|---|---|---|---|---|---|
| Mess-zeitpunkt | $\chi^2$ | $p\,(\chi^2)$ | $\chi^2/df$ | CFI | RMSEA | SRMR |
| $t_1$ | 783.44 | < .01 | 2.73 | .92 | 0.050 | 0.055 |
| $t_2$ | 536.21 | < .01 | 2.74 | .93 | 0.050 | 0.053 |
| $t_3$ | 816.45 | < .01 | 2.84 | .92 | 0.051 | 0.060 |
| $t_4$ | 507.85 | < .01 | 2.59 | .93 | 0.048 | 0.051 |

*Anmerkungen.* $t_1$: während des Tests (1. Teststunde), $t_2$: nach dem Test (1. Teststunde), $t_3$: während des Tests (2. Teststunde), $t_4$: während des Tests (2. Teststunde).

Hypothese 2 nimmt einen positiven Zusammenhang zwischen der Erwartungs- und der Wertkomponente des Modells an. Zunächst sollen diese beiden Komponenten deskriptiv betrachtet werden. Dazu sind in Tabelle 7.4 die Mittelwerte und Standardabweichungen der beiden Variablen für die vier Messzeitpunkte dargestellt. Die Erwartung einer erfolgreichen Testbearbeitung liegt zu Beginn der Testung in mittlerer Höhe. Innerhalb der Teststunden nimmt die Erwartung jeweils im Verlauf der Testbearbeitung numerisch ab. In der ersten Teststunde ist die Erwartung insgesamt höher ausgeprägt als in der zweiten Teststunde. Der Wert einer erfolgreichen Testbearbeitung befindet sich in der ersten Teststunde über dem Skalenmittelwert, in der zweiten Teststunde hingegen nur noch in mittlerer Höhe. Innerhalb der Teststunden ist kein Absinken des Werts zu erkennen.

*Tabelle 7.4*: Deskriptive Angaben zur Erwartungs- und Wertkomponente der Motivation zur Testbearbeitung.

| | $t_1$ | | $t_2$ | | $t_3$ | | $t_4$ | |
|---|---|---|---|---|---|---|---|---|
| Variable | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Erwartung einer erfolgreichen Testbearbeitung[a] | 2.49 | 0.55 | 2.35 | 0.66 | 2.43 | 0.64 | 2.28 | 0.72 |
| Wert einer erfolgreichen Testbearbeitung[b] | 2.71 | 0.64 | 3.15 | 0.80 | 2.48 | 0.75 | 3.06 | 0.81 |

*Anmerkungen.* [a] Antwortskala für $t_1$ bis $t_4$ von 1 bis 4; [b] Antwortskala für $t_1$ und $t_3$ von 1 bis 4, für $t_2$ und $t_4$ von 1 bis 5 (vgl. Abschnitt 6.2.2); $t_1$: während des Tests (1. Teststunde), $t_2$: nach dem Test (1. Teststunde), $t_3$: während des Tests (2. Teststunde), $t_4$: während des Tests (2. Teststunde).

Wie in Abbildung 7.3 und in Abbildung 11.3 bis Abbildung 11.6 im Anhang zu sehen, kann die Annahme eines positiven Zusammenhangs zwischen der Erwartungs- und der Wertkomponente (vgl. Hypothese 2) für den ersten und den dritten Messzeitpunkt mit Effekten in kleiner bis mittlerer Höhe bestätigt werden ($\psi^{S}_{21,t1}$ = .25, $p$ < .01; $\psi^{S}_{21,t3}$ = .32, $p$ < .01). Auch für die anderen beiden Messzeitpunkte zeigen sich positive Zusammenhänge zwischen den beiden Variablen, die aber statistisch nicht gegen den Zufall abgesichert werden können ($\psi^{S}_{21,t2}$ = .14, $p$ = .06; $\psi^{S}_{21,t4}$ = .11, $p$ = .14). Hypothese 2 kann demnach lediglich für die Motivation zur Testbearbeitung während eines Kompetenztests bestätigt werden.

Mit Hypothese 3 wird angenommen, dass die Erwartungskomponente für die Vorhersage der Testleistung in einem Test ohne Aufgabenwahlmöglichkeit wichtiger ist als die Wertkomponente. Um die Gültigkeit dieser Hypothese zu prüfen, werden die Pfadkoeffizienten zwischen der Erwartung und der Testleistung ($\beta^{S}_{Y1}$) sowie zwischen dem Wert und der Testleistung ($\beta^{S}_{Y2}$) vergleichend betrachtet. Der prädiktive Effekt der Erwartung auf die Testleistung ist numerisch zu allen vier Messzeitpunkten größer als der des Werts (Range der Koeffizienten für die Erwartung: .29 bis .36; Range der Koeffizienten für den Wert: -.10 bis .07; vgl. Abbildung 7.3 sowie Abbildung 11.3 bis Abbildung 11.6 im Anhang). Ein inferenzstatistischer Vergleich der Pfadkoeffizienten bestätigt, dass die Erwartung für die Vorhersage der Testleistung bedeutsamer ist als der Wert ($\Delta\beta_{Y1\text{-}Y2,t1}$ = .45, $p$ < .01; $\Delta\beta_{Y1\text{-}Y2,t2}$ = .41, $p$ < .01; $\Delta\beta_{Y1\text{-}Y2,t3}$ = .43, $p$ < .01; $\Delta\beta_{Y1\text{-}Y2,t4}$ = .27, $p$ < .01). Darüber hinaus lässt sich der positive Einfluss der Erwartung auf die Testleistung zu allen vier Messzeitpunkten gegenüber dem Zufall absichern (für alle Messzeitpunkte: $p$ < .01), während der Wert lediglich zum zweiten Messzeitpunkt

eine signifikant negative Wirkung auf die Testleistung hat ($\beta^{s}_{Y2,t2}$ = -.10, $p$ < .05). Generell gilt demnach, dass die Testleistung umso höher ist, je größer die Erfolgserwartung der Testpersonen ist. Nach Ende des ersten Tests schneiden die Testpersonen dagegen umso schlechter im Kompetenztest ab, je wichtiger ihnen die erfolgreiche Testbearbeitung ist. Dieser erwartungswidrige Effekt ist allerdings klein und zeigt sich nur zu diesem einen Messzeitpunkt; für alle anderen Messzeitpunkte ist der Wert einer erfolgreichen Testbearbeitung für die aktuelle Testleistung irrelevant. Es ist zusammenzufassen, dass die Erwartung für die Vorhersage der Testleistung wichtiger ist als der Wert. Die Erwartung sagt die Testleistung statistisch signifikant und positiv vorher, während der Wert für die Testleistung kaum von Bedeutung ist (weshalb er wie vermutet in den Hypothesentests zur zweiten Fragestellung nur am Rande Beachtung finden wird). Hypothese 3 gilt als bestätigt.

Hypothese 4 behauptet, dass das Erwartung-Wert-Modell der Motivation zur Testbearbeitung besser passt, wenn die State- und die Trait-Leistungsmotivation gleichzeitig berücksichtigt werden als wenn lediglich die State-Leistungsmotivation betrachtet wird. Zur Prüfung dieser Hypothese wird ein $\chi^2$-Differenzentest mit Satorra-Bentler-Korrektur durchgeführt. Dieser Test vergleicht die Anpassung des unrestringierten Modells, in welchem der Einfluss der Trait-Leistungsmotivation (Hoffnung auf Erfolg; für deskriptive Werte siehe Tabelle 6.2) frei geschätzt wird, mit der Anpassung eines restringierten Modells, in welchem der Einfluss der Trait-Leistungsmotivation auf Null fixiert wird. Die Prüfgröße TRd errechnet sich dabei folgendermaßen (vgl. http://www.statmodel.com/chidiff.shtml; Satorra & Bentler, 1999):

$$\mathrm{TRd} = \frac{(\chi_0^2 \cdot c_0 - \chi_1^2 \cdot c_1) \cdot (df_0 - df_1)}{(df_0 \cdot c_0 - df_1 \cdot c_1)}$$

wobei
- $\chi_0^2$ Chi-Quadrat-Wert des restringierten Modells
- $\chi_1^2$ Chi-Quadrat-Wert des unrestringierten Modells
- $c_0$ Korrekturfaktor des restringierten Modells
- $c_1$ Korrekturfaktor des unrestringierten Modells
- $df_0$ Freiheitsgrade des restringierten Modells
- $df_1$ Freiheitsgrade des unrestringierten Modells

Zu allen vier Messzeitpunkten ergibt sich eine signifikant bessere Modellanpassung des ursprünglichen, unrestringierten Modells gegenüber dem restringierten Modell ($\Delta\chi^2_{t1}$ = 55.83, $\Delta df$ = 1, $\mathrm{TRd}_{t1}$ = 45.10, $p$ < .01; $\Delta\chi^2_{t2}$ = 71.48, $\Delta df$ = 1, $\mathrm{TRd}_{t2}$ = 35.85, $p$ < .01; $\Delta\chi^2_{t3}$ = 66.10, $\Delta df$ = 1, $\mathrm{TRd}_{t3}$ = 44.83, $p$ < .01; $\Delta\chi^2_{t4}$ = 96.59, $\Delta df$ = 1, $\mathrm{TRd}_{t4}$ = 82.69, $p$ < .01). Dieses Ergebnis überrascht nicht, betrachtet man die Höhe der Pfadkoeffizienten zwischen Hoffnung auf Erfolg und Wert ($\gamma_{23}$), die zwischen .36 und .52 liegen und alle auf dem 1-Prozent-Niveau statistisch signifikant sind. Je stärker das Leistungsmotiv Hoffnung auf Erfolg ausgeprägt ist, umso stärker wird eine erfolgreiche Testbearbeitung wertgeschätzt. Das Leistungsmotiv spielt demnach für die aktuelle Motivation zur Testbearbeitung eine wichtige Rolle. Hypothese 4 gilt als bestätigt.

In Ergänzung zu den Ergebnissen der Hypothesentests zur ersten Fragestellung wird im Folgenden auf einige weitere Befunde aus dem Erwartung-Wert-Modell der Motivation zur Testbearbeitung hingewiesen: Wie theoretisch zu erwarten, zeigen sich empirisch die größten positiven Zusammenhänge zwischen der wahrgenommenen Testleistung und der Erwartung, zwischen der Erwartung und der tatsächlichen Testleistung sowie zwischen der Hoffnung auf Erfolg

und dem Wert. Das Zusammenhangsmuster zwischen den Variablen des Modells ist über die Zeit hinweg recht stabil (vgl. Abbildung 11.3 bis Abbildung 11.6 im Anhang), was die Modellgeltung über die genannten Fit-Kriterien hinaus untermauert.

## 7.3 Eignung der Daten für Messwiederholungsanalysen

Bevor die Hypothesen zur zweiten Fragestellung anhand von Mehrebenen-Wachstumsmodellen überprüft werden können, muss sichergestellt werden, dass die Voraussetzungen für die Analyse von Messwiederholungsdaten gegeben sind. So muss das Vorhandensein von Carryover-Effekten ausgeschlossen werden, da solche Effekte die Ergebnisse verzerren würden (Abschnitt 7.3.1). Zudem muss gewährleistet sein, dass hinsichtlich der abhängigen Variablen Erwartung und Wert einer erfolgreichen Testbearbeitung mindestens starke faktorielle Invarianz vorliegt (Abschnitt 7.3.2).

### 7.3.1 Carryover-Effekte

Um die Daten auf Carryover-Effekte zu überprüfen, werden Multigruppenanalysen durchgeführt. Diese Analysen dienen dazu, die Veränderung der Motivation zur Testbearbeitung von der ersten zur zweiten Teststunde zwischen den beiden Gruppen „FIT – CAT“ und „CAT – FIT“ zu vergleichen. Es wird geprüft, ob sich die Motivation in Abhängigkeit davon, welcher Testalgorithmus zuerst dargeboten wurde, zwischen der ersten und der zweiten Teststunde unterschiedlich verändert. Die Multigruppenanalysen werden separat für den Zeitpunkt „während des Tests“ und für den Zeitpunkt „nach dem Test“ durchgeführt, um lediglich identische Motivationsskalen miteinander zu vergleichen. Die Analysen erfolgen sowohl für die Erwartungs- als auch für die Wertkomponente und unter Berücksichtigung der komplexen Stichprobenstruktur. Zunächst wird jeweils innerhalb der beiden Gruppen die Veränderung der Motivation zur Testbearbeitung von der ersten zur zweiten Teststunde berechnet und gegen Null getestet. Anschließend wird über eine Differenzbildung der Veränderungswerte aus den beiden Gruppen geprüft, ob sich die Motivation zwischen den beiden Teststunden je nach Gruppenzugehörigkeit unterschiedlich verändert. Falls die Differenz der Veränderungen signifikant ausfällt, muss dies als Hinweis auf das Vorliegen von Carryover-Effekten gedeutet werden. Im Folgenden werden die Ergebnisse der Multigruppenanalysen der Reihe nach für die Erwartung während des Tests, für die Erwartung nach dem Test, für den Wert während des Tests und für den Wert nach dem Test dargestellt.

*Tabelle 7.5*: Ergebnisse der Multigruppenanalyse für die Erwartung während des Tests.

| | $t_1$ | | $t_3$ | | $\Delta M_{t3\text{-}t1}$ | | | $\Delta(\Delta M_{t3\text{-}t1})_{FIT\text{-}CAT\ –\ CAT\text{-}FIT}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gruppe | *M* | *SD* | *M* | *SD* | *M* | *SE* | *p*(Δ) | *M* | *SE* | *p*(Δ(Δ)) |
| FIT-CAT | 2.51 | 0.57 | 2.46 | 0.64 | -0.05 | 0.02 | < .01 | | | |
| | | | | | | | | 0.02 | 0.03 | .39 |
| CAT-FIT | 2.48 | 0.54 | 2.40 | 0.65 | -0.08 | 0.03 | < .01 | | | |

*Anmerkungen.* $t_1$: während des Tests (1. Teststunde), $t_3$: während des Tests (2. Teststunde); FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

Tabelle 7.5 zeigt die gruppenspezifischen Mittelwerte der Erwartung während des Tests in der ersten und zweiten Teststunde ($t_1$ bzw. $t_3$). Numerisch besteht in der Gruppe, die zuerst den FIT und dann den CAT bearbeitet hat, eine etwas höhere Erwartung einer erfolgreichen Testbearbeitung als in der Gruppe, die zuerst den CAT und dann den FIT bearbeitet hat. In beiden Gruppen nimmt die Erwartung von der ersten zur zweiten Teststunde jedoch signifikant ab ($\Delta M_{t3\text{-}t1}$). Dies ist im Sinne eines einfachen Reihenfolgeeffekts zu interpretieren, der möglicherweise auf Ermüdung zurückzuführen ist. Dieser Effekt wirkt sich gleichermaßen auf beide Testbedingungen aus und ist für die Interpretation der UV-Effekte daher unerheblich. Carryover-Effekte, also unterschiedlich starke Veränderungen der Motivation in Abhängigkeit von der Reihenfolge der Testalgorithmen, sind für die Erwartungskomponente während des Tests nicht festzustellen ($\Delta(\Delta M_{t3\text{-}t1})_{FIT\text{-}CAT\ -\ CAT\text{-}FIT}$). Numerisch sinkt die Motivation zwar geringfügig mehr, wenn zuerst der CAT und dann der FIT vorgegeben wird, als wenn die Tests in umgekehrter Reihenfolge bearbeitet werden, doch ist diese Differenz zwischen den Veränderungen nicht signifikant.

*Tabelle 7.6*: Ergebnisse der Multigruppenanalyse für die Erwartung nach dem Test.

| | $t_2$ | | $t_4$ | | $\Delta M_{t4\text{-}t2}$ | | | $\Delta(\Delta M_{t4\text{-}t2})_{FIT\text{-}CAT\ -\ CAT\text{-}FIT}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gruppe | *M* | *SD* | *M* | *SD* | *M* | *SE* | *p*(Δ) | *M* | *SE* | *p*(Δ(Δ)) |
| FIT-CAT | 2.36 | 0.67 | 2.32 | 0.75 | -0.04 | 0.03 | .24 | | | |
| | | | | | | | | 0.05 | 0.05 | .33 |
| CAT-FIT | 2.34 | 0.65 | 2.25 | 0.68 | -0.08 | 0.04 | < .05 | | | |

*Anmerkungen.* $t_2$: nach dem Test (1. Teststunde), $t_4$: nach dem Test (2. Teststunde); FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

Ähnliche Ergebnisse finden sich bei den übrigen Multigruppenanalysen (Tabelle 7.6 bis Tabelle 7.8): So zeigt sich durchweg eine numerisch höhere Motivation zur Testbearbeitung in der Gruppe, die zuerst den FIT und dann den CAT bearbeitet hat. Generell nimmt die Motivation von der ersten zur zweiten Teststunde hin signifikant ab. Lediglich die Erwartung einer erfolgreichen Testbearbeitung nach dem Test bleibt in der Gruppe FIT-CAT konstant (Tabelle 7.6).

*Tabelle 7.7*: Ergebnisse der Multigruppenanalyse für den Wert während des Tests.

| | $t_1$ | | $t_3$ | | $\Delta M_{t3\text{-}t1}$ | | | $\Delta(\Delta M_{t3\text{-}t1})_{FIT\text{-}CAT\ -\ CAT\text{-}FIT}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gruppe | *M* | *SD* | *M* | *SD* | *M* | *SE* | *p*(Δ) | *M* | *SE* | *p*(Δ(Δ)) |
| FIT-CAT | 2.74 | 0.65 | 2.51 | 0.77 | -0.22 | 0.04 | < .01 | | | |
| | | | | | | | | 0.00 | 0.05 | .96 |
| CAT-FIT | 2.68 | 0.62 | 2.45 | 0.72 | -0.23 | 0.03 | < .01 | | | |

*Anmerkungen.* $t_1$: während des Tests (1. Teststunde), $t_3$: während des Tests (2. Teststunde); FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

Numerisch zeigt sich generell eine stärkere Motivationsabnahme in der Gruppe, die zuerst den CAT und dann den FIT bearbeitet hat, als in der Gruppe, die erst den FIT und dann den CAT vorgegeben bekommen hat. Das heißt, die Gruppe CAT-FIT bescheinigt sich selbst nicht nur generell eine

numerisch geringere Motivation als die Gruppe FIT-CAT, sondern sie berichtet darüber hinaus auch einen stärkeren Motivationsverlust. Die Differenz der Motivationsabnahme zwischen den beiden Gruppen ist jedoch nicht nur für die Erwartung während des Tests, sondern auch für die Erwartung nach dem Test sowie für den Wert während des Tests und nach dem Test statistisch nicht signifikant. Für die vorliegende Studie liegen demnach keine Carryover-Effekte vor.

*Tabelle 7.8*: Ergebnisse der Multigruppenanalyse für den Wert nach dem Test.

| | $t_2$ | | $t_4$ | | $\Delta_{t4\text{-}t2}$ | | | $\Delta(\Delta_{t4\text{-}t2})_{FIT\text{-}CAT\ –\ CAT\text{-}FIT}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gruppe | *M* | *SD* | *M* | *SD* | *M* | *SE* | *p*(Δ) | *M* | *SE* | *p*(Δ(Δ)) |
| FIT-CAT | 3.19 | 0.85 | 3.12 | 0.84 | -0.07 | 0.03 | < .05 | | | |
| | | | | | | | | 0.03 | 0.04 | .48 |
| CAT-FIT | 3.11 | 0.75 | 3.01 | 0.77 | -0.10 | 0.03 | < .01 | | | |

*Anmerkungen.* $t_2$: nach dem Test (1. Teststunde), $t_4$: nach dem Test (2. Teststunde); FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

### 7.3.2 Faktorielle Invarianz

Neben dem Ausschluss von Carryover-Effekten muss als weitere Voraussetzung für die Analyse der Messwiederholungsdaten im Rahmen von Mehrebenen-Wachstumsmodellen geprüft werden, ob mindestens starke faktorielle Invarianz zwischen den Daten der ersten und der zweiten Teststunde vorliegt. Diese Prüfung erfolgt über Modellvergleiche zwischen unrestringierten Modellen, in denen keinerlei Invarianzanforderungen für die beiden Messzeitpunkte definiert werden, und restringierten Modellen, die Invarianzanforderungen unterliegen. Im unrestringierten Modell dürfen sich die Schätzungen der Modellparameter zwischen der ersten und der zweiten Teststunde somit unterscheiden, während sie im restringierten Modell Gleichheitsrestriktionen unterworfen sind. Welche Parameter im Rahmen der Prüfung der faktoriellen Invarianz gleichgesetzt werden, wird exemplarisch in Bezug auf die Erwartungskomponente der Motivation während des Tests verdeutlicht (Abbildung 7.4).

Wie in Abbildung 7.4 zu sehen, beruht die (latente) Erwartungskomponente auf fünf Indikatoren, die während des Tests in der ersten Teststunde und während des Tests in der zweiten Teststunde per Fragebogen erfasst wurden. Da die Fragebogenskala zu beiden Messzeitpunkten dieselben Fragen enthält, wird die Autokorrelation zwischen den Residuen korrespondierender Indikatoren zugelassen (rekursive Pfeile zwischen den Residualvariablen $\varepsilon_{ik}$). Der Fehlerterm der ersten Frage aus der ersten Teststunde darf also mit dem Fehlerterm der ersten Frage aus der zweiten Teststunde korrelieren. Im unrestringierten Modell wird lediglich die Faktorstruktur vorgegeben (d. h. die latente Variable Erwartung beruht auf fünf Indikatoren). Die Modellparameter werden für beide Messzeitpunkte frei geschätzt. Im restringierten Modell mit starker faktorieller Invarianz werden die Faktorladungen und die Intercepts korrespondierender Indikatoren zwischen den beiden Messzeitpunkten gleichgesetzt (d. h. $\lambda_{1k} = \lambda_{3k}$ und $\alpha_{1k} = \alpha_{3k}$). Im restringierten Modell mit strikter faktorieller Invarianz werden zusätzlich die Messfehlervarianzen gleichgesetzt (d. h. $Var(\varepsilon_{1k}) = Var(\varepsilon_{3k})$; vgl. Geiser 2010). Im Folgenden werden die Ergebnisse der Invarianztests für die Erwartungs- und die Wertkomponente während des Tests und nach dem Test berichtet.

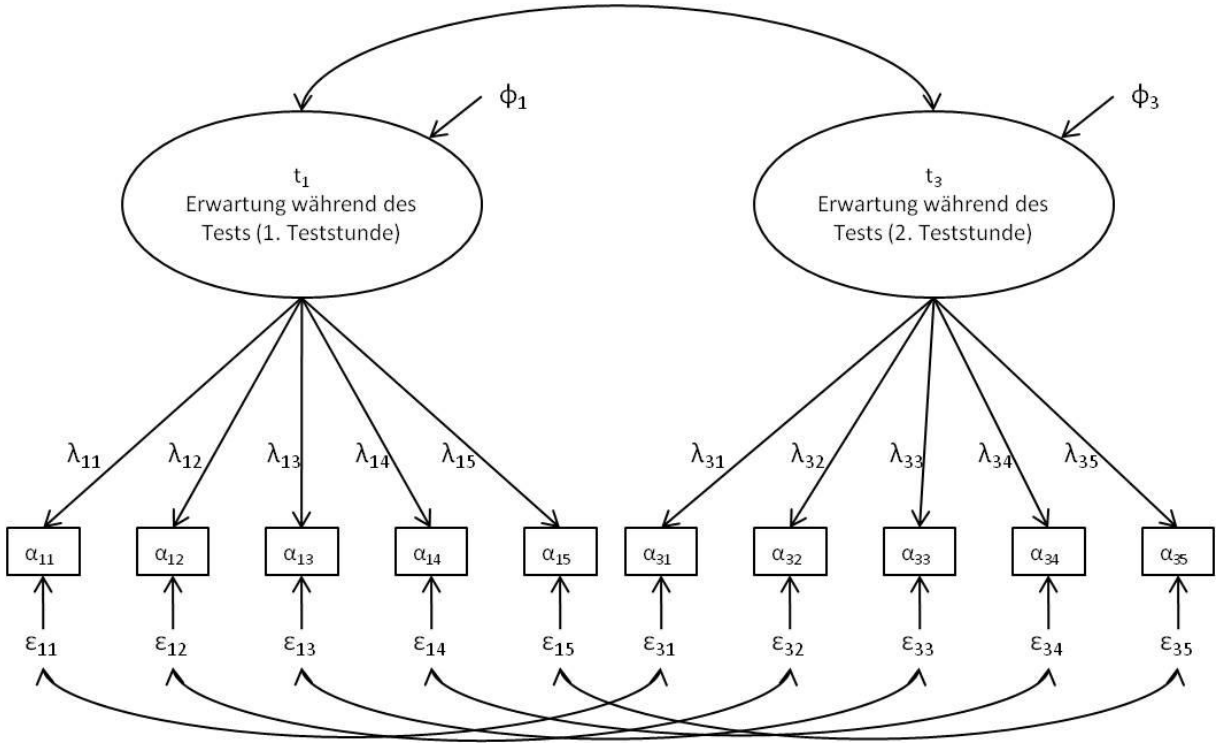



*Abbildung 7.4*: Modell zur Prüfung der faktoriellen Invarianz der Erwartung einer erfolgreichen Testbearbeitung während des Tests zwischen der ersten und der zweiten Teststunde ($\phi_i$: Residualvarianz der latenten Variable zum Zeitpunkt *i*, $\lambda_{ik}$: Faktorladung des Indikators *k* zum Zeitpunkt *i*, $\alpha_{ik}$: Intercept des Indikators *k* zum Zeitpunkt *i*, $\varepsilon_{ik}$: Residualvariable des Indikators *k* zum Zeitpunkt *i*).

Für die Erwartungskomponente liegt sowohl während des Tests als auch nach dem Test strikte faktorielle Invarianz vor. Somit erweisen sich nicht nur die Faktorenstruktur, die Faktorladungen und die Intercepts, sondern auch die Messfehlervarianzen zwischen den beiden Messzeitpunkten als vergleichbar (Tabelle 7.9; für Erläuterungen der einzelnen Messinvarianzbedingungen sei auf Geiser, 2010 verwiesen).

*Tabelle 7.9:* Anpassungskriterien verschiedener Messinvarianzbedingungen für die Erwartungskomponente.

| Messinvarianz-bedingung | $\chi^2$ | *df* | $\Delta\chi^2$ | $\Delta df$ | *p* ($\Delta\chi^2$) |
|---|---|---|---|---|---|
| $t_1$-$t_3$ unrestringiert | 54.04 | 29 | | | |
| | | | 12.19 | 13 | .51 |
| $t_1$-$t_3$ strikt | 66.23 | 42 | | | |
| $t_2$-$t_4$ unrestringiert | 24.90 | 1 | | | |
| | | | 0.06 | 3 | > .99 |
| $t_2$-$t_4$ strikt | 24.84 | 4 | | | |

*Anmerkungen.* $t_1$: während des Tests (1. Teststunde), $t_2$: nach dem Test (1. Teststunde), $t_3$: während des Tests (2. Teststunde), $t_4$: während des Tests (2. Teststunde).

Für die Wertkomponente hingegen kann weder während des Tests noch nach dem Test die Mindestanforderung starker faktorieller Invarianz bestätigt werden (Tabelle 7.10). Somit ist die sinnvolle Durchführung von Messwiederholungsanalysen anhand der vorliegenden Daten lediglich für die Erwartungskomponente möglich. Daher beschränken sich die folgenden Hypothesentests auf diese Komponente der aktuellen Motivation zur Testbearbeitung als abhängige Variable. Da theoretisch ohnehin kaum Effekte auf die Wertkomponente zu erwarten sind (vgl. Abschnitt 5) und die Wertkomponente kaum in Zusammenhang mit der aktuellen Testleistung steht, ist diese Einschränkung unerheblich (vgl. Abschnitt 7.2).

*Tabelle 7.10:* Anpassungskriterien verschiedener Messinvarianzbedingungen für die Wertkomponente.

| Messinvarianz-bedingung | $\chi^2$ | *df* | $\Delta\chi^2$ | $\Delta df$ | $p$ ($\Delta\chi^2$) |
|---|---|---|---|---|---|
| $t_1$-$t_3$ unrestringiert | 73.38 | 29 | | | |
| | | | 54.32 | 8 | < .01 |
| $t_1$-$t_3$ stark | 127.70 | 37 | | | |
| $t_2$-$t_4$ unrestringiert | 55.71 | 15 | | | |
| | | | 37.15 | 6 | < .01 |
| $t_2$-$t_4$ stark | 92.85 | 21 | | | |

*Anmerkungen.* $t_1$: während des Tests (1. Teststunde), $t_2$: nach dem Test (1. Teststunde), $t_3$: während des Tests (2. Teststunde), $t_4$: während des Tests (2. Teststunde).

## 7.4 Effekte von Testalgorithmus und Selbstkonzept auf die Motivation zur Testbearbeitung

Für die Daten der Erwartungskomponente sind die Voraussetzungen für Messwiederholungsanalysen gegeben. Diese Daten können somit für die erste und zweite Teststunde gemeinsam analysiert werden, so dass die volle Stichprobengröße ausgeschöpft wird. Daher werden die Hypothesen der zweiten Fragestellung anhand von Mehrebenen-Wachstumsmodellen überprüft, in die der Messzeitpunkt (erste oder zweite Teststunde) auf Ebene 1 und die Testpersonen auf Ebene 2 eingehen. Die komplexe Stichprobenstruktur wird über die Stratifizierungsvariable „Schulart“ und die Clustervariable „Klasse“ abgebildet. Zur Prüfung der Hypothesen 5 und 6 werden auf Ebene 1 die beiden Haupteffekte Messzeitpunkt und Testalgorithmus modelliert, auf Ebene 2 (unter Kontrolle von Hoffnung auf Erfolg) der Haupteffekt des Fähigkeitsselbstkonzepts sowie die Zwischen-Ebenen-Interaktionseffekte Testalgorithmus x Fähigkeitsselbstkonzept und Testalgorithmus x Hoffnung auf Erfolg. Das Fähigkeitsselbstkonzept und die Hoffnung auf Erfolg werden an ihrem Mittelwert zentriert (*grand mean centering*), um die Interpretation der Ergebnisse zu erleichtern (vgl. Luke, 2004). Für ein besseres Verständnis des verwendeten Mehrebenen-Wachstumsmodells wird im Folgenden das Gleichungssystem der beiden Ebenen aufgeführt:

Ebene 1: $y_{ij} = \beta_{0j} + \beta_{1j}(MZP) + \beta_{2j}(TA) + r_{ij}$

Ebene 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(SK) + \gamma_{02}(H) + u_{0j}$
$\beta_{1j} = \gamma_{10} + u_{1j}$
$\beta_{2j} = \gamma_{20} + \gamma_{21}(SK) + \gamma_{22}(H) + u_{2j}$

wobei

| | |
|---|---|
| *y* | Erfolgserwartung |
| *i* | Indikator für Messzeitpunkt |
| *j* | Indikator für Testperson |
| MZP | Messzeitpunkt |
| TA | Testalgorithmus (0: FIT, 1: CAT) |
| SK | Fähigkeitsselbstkonzept Mathematik |
| H | Hoffnung auf Erfolg |

Das Gleichungssystem macht deutlich, dass für jede Testperson ein eigenes Ebene-1-Modell (Wachstumsmodell) geschätzt wird. Für das Intercept der Erfolgserwartung ($\beta_{0j}$), also die absolute Höhe, werden die Haupteffekte von Fähigkeitsselbstkonzept und Hoffnung auf Erfolg als feste Effekte auf Ebene 2 modelliert. Für den Einfluss des Messzeitpunkts auf die Erfolgserwartung ($\beta_{1j}$) werden keine differentiellen Effekte in Abhängigkeit bestimmter Personenvariablen berücksichtigt, da hierzu keine theoretischen Annahmen vorliegen. Für den Einfluss des Testalgorithmus auf die Erfolgserwartung ($\beta_{2j}$) hingegen werden, in Übereinstimmung mit den formulierten Hypothesen, feste Interaktionseffekte mit den Ebene-2-Variablen Fähigkeitsselbstkonzept und Hoffnung auf Erfolg modelliert. Darüber hinaus werden die Varianzen auf Ebene 1 und Ebene 2 als zufällige Effekte berücksichtigt. Exemplarisch wird im Text im Folgenden jeweils die Tabelle mit den Ergebnissen für die Gesamtstichprobe präsentiert. Im Text werden lediglich jene Koeffizienten genannt, die für die Prüfung der Hypothesen relevant sind. Die Ergebnisse aller durchgeführten Mehrebenen-Wachstumsmodelle sind dem Anhang zu entnehmen (Tabelle 11.1 bis Tabelle 11.6).

Die Prüfung der Hypothesen 5 und 6 erfolgt jeweils zunächst unter Betrachtung der Gesamtstichprobe. Anschließend werden explorativ separate Analysen für Personen mit hoher beziehungsweise niedriger Hoffnung auf Erfolg sowie für Personen aus den Schularten Hauptschule ($n$ = 140), Realschule ($n$ = 270) und Gymnasium ($n$ = 251) durchgeführt. Die Einteilung in die beiden Leistungsmotiv-Kategorien erfolgt kriterial, indem Personen mit einem Skalenmittelwert von ≤ 2 eine geringe Hoffnung auf Erfolg ($n$ = 92) und Personen mit einem Skalenmittelwert ≥ 3 eine hohe Hoffnung auf Erfolg ($n$ = 294) zugeschrieben wird. Die Schulart Integrierte Gesamtschule wird von den separaten Betrachtungen ausgenommen, da die Zellenbesetzungen in dieser Schulart zu klein sind ($n$ = 42). Tabelle 7.11 gibt einen deskriptiven Überblick über die Ausprägungen der Erfolgserwartung im FIT und im CAT für die Gesamtstichprobe und für die explorativ untersuchten Substichproben.

*Tabelle 7.11*: Mittelwerte und Standardabweichungen der Erfolgserwartung im FIT und im CAT.

| | | Während des Tests | | Nach dem Test | |
|---|---|---|---|---|---|
| Stichprobe | Testalgo-rithmus | *M* | *SD* | *M* | *SD* |
| Gesamtstichprobe | | | | | |
| | FIT | 2.46 | 0.61 | 2.30 | 0.68 |
| | CAT | 2.47 | 0.59 | 2.33 | 0.70 |
| Personen mit hoher Hoffnung auf Erfolg | | | | | |
| | FIT | 2.66 | 0.60 | 2.50 | 0.69 |
| | CAT | 2.64 | 0.59 | 2.50 | 0.70 |
| Personen mit niedriger Hoffnung auf Erfolg | | | | | |
| | FIT | 2.05 | 0.61 | 1.90 | 0.63 |
| | CAT | 2.12 | 0.65 | 2.00 | 0.76 |
| Hauptschülerinnen/-schüler | | | | | |
| | FIT | 2.42 | 0.62 | 2.31 | 0.69 |
| | CAT | 2.43 | 0.65 | 2.32 | 0.75 |
| Realschülerinnen/-schüler | | | | | |
| | FIT | 2.43 | 0.60 | 2.21 | 0.67 |
| | CAT | 2.43 | 0.57 | 2.28 | 0.69 |
| Gymnasiastinnen/Gymnasiasten | | | | | |
| | FIT | 2.54 | 0.60 | 2.44 | 0.65 |
| | CAT | 2.56 | 0.56 | 2.43 | 0.67 |

*Anmerkungen.* FIT: computerisierter nicht-adaptiver Test; CAT: computerisierter adaptiver Test.

An den in Tabelle 7.12 dargestellten Ergebnissen zum Mehrebenen-Wachstumsmodell für die Gesamtstichprobe ist zunächst der positive Effekt des Fähigkeitsselbstkonzepts auf die Erfolgserwartung wiederzuerkennen ($\gamma_{01}$), der bereits in den Erwartung-Wert-Modellen der Motivation zur Testbearbeitung gezeigt wurde (Abschnitt 7.2). Auch die Hoffnung auf Erfolg steht in einem positiven Zusammenhang zur Erfolgserwartung ($\gamma_{02}$). Darüber hinaus findet sich zudem die Abnahme der Erfolgserwartung von der ersten zur zweiten Teststunde wieder ($\gamma_{10}$), die in Abschnitt 7.2 bereits deskriptiv festgestellt wurde. Ebenfalls wie erwartet, fällt der ICC, also der Anteil der Varianz der Erfolgserwartung, der auf Merkmale der Personen (Ebene 2) zurückzuführen ist, mit .73 während des Tests und .59 nach dem Test sehr hoch aus. Andersherum betrachtet, können 27 Prozent beziehungsweise 41 Prozent der Varianz in der Erfolgserwartung durch Einflüsse des Messzeitpunkts (Ebene 1) erklärt werden. Die proportionale Verringerung der Residualvarianz des Nullmodells durch das restringierte Mehrebenen-Wachstumsmodell (vgl. Erläuterungen zur Modellgüte in Abschnitt 6.5.4.2) von $R^2$ = .38 während des Tests und von $R^2$ = .24 nach dem Test liegt in moderater Höhe. Es soll jedoch in Erinnerung gerufen werden, dass diese Werte die Modellgüte des vollständigen, unrestringierten Modells vermutlich unterschätzen. Denn die $R^2$-Werte mussten anhand eines restringierten Modells berechnet werden, das keine zufälligen Effekte auf Ebene 2 enthält. Angesichts der sehr hohen ICCs ist davon auszugehen, dass die Modellgüte des restringierten Modells durch diese massive Restriktion deutlich schlechter ausfällt als die des vollständigen Modells.

Hypothese 5 postuliert, dass es keinen Haupteffekt des Testalgorithmus ($\gamma_{20}$) auf die Motivation zur Testbearbeitung gibt. Wie erwartet, wird der Haupteffekt weder während des Tests ($\gamma_{20} = 0.00$, $p = .89$) noch nach dem Test signifikant ($\gamma_{20} = 0.02$, $p = .43$; Tabelle 7.12). Die Erfolgserwartung unterscheidet sich also erwartungsgemäß im FIT und im CAT über alle Personen hinweg nicht. Auch die separate Betrachtung der Personen mit hoher und mit niedriger Hoffnung auf Erfolg und die separate Betrachtung der Schularten erbringen keinen signifikanten Haupteffekt des Testalgorithmus (vgl. Tabelle 11.2 bis Tabelle 11.6 im Anhang). Es ist festzuhalten, dass sich die mittlere Erfolgserwartung im FIT erwartungsgemäß nicht von der mittleren Erfolgserwartung im CAT unterscheidet. Hypothese 5 wird bestätigt.

*Tabelle 7.12*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.49 | 0.02 | 154.78 | < .01 | 2.34 | 0.02 | 111.56 | < .01 |
| SK ($\gamma_{01}$) | 0.41 | 0.03 | 11.73 | < .01 | 0.32 | 0.04 | 7.49 | < .01 |
| H ($\gamma_{02}$) | 0.25 | 0.03 | 8.73 | < .01 | 0.28 | 0.04 | 6.88 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.02 | -3.40 | < .01 | -0.06 | 0.03 | -2.40 | < .05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.00 | 0.01 | 0.14 | .89 | 0.02 | 0.02 | 0.80 | .43 |
| SK ($\gamma_{21}$) | -0.02 | 0.02 | -0.79 | .43 | 0.01 | 0.03 | 0.16 | .87 |
| H ($\gamma_{22}$) | -0.04 | 0.03 | -1.13 | .26 | -0.05 | 0.04 | -1.09 | .28 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .38; nach dem Test: *ICC* = .59; $R^2$ = .24;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

In Hypothese 6 wird ein Zwischen-Ebenen-Interaktionseffekt von Testalgorithmus und Fähigkeitsselbstkonzept ($\gamma_{21}$) auf die Motivation zur Testbearbeitung angenommen. Wider Erwarten wird dieser Interaktionseffekt jedoch in der Gesamtstichprobe weder während des Tests ($\gamma_{21} = -0.02$, $p = .43$) noch nach dem Test signifikant ($\gamma_{21} = 0.01$, $p = .87$; Tabelle 7.12).

Bei separater Analyse der beiden Leistungsmotivgruppen hingegen zeigt sich der Interaktionseffekt von Testalgorithmus und Fähigkeitsselbstkonzept zumindest für Personen mit hoher Hoffnung auf Erfolg nach dem Test ($\gamma_{21} = -0.08$, $p < .05$; Tabelle 11.2). Simple-Slope-Analysen eröffnen, dass dieser Effekt auf diejenigen Personen mit hoher Hoffnung auf Erfolg zurückgeht, die über ein hohes Fähigkeitsselbstkonzept verfügen: Diese Personen sind im CAT signifikant geringer motiviert als im FIT ($\gamma_{20} = -0.06$, $p < .05$; bestätigt Hypothese 6a). Besteht zwar eine hohe dispositionelle Hoffnung auf Erfolg, aber ein niedriges Fähigkeitsselbstkonzept, besteht zwischen CAT und FIT kein Motivationsunterschied ($\gamma_{20} = 0.06$, $p = .29$; widerspricht Hypothese 6b).

Die Betrachtung der Schularten eröffnet ebenfalls ein interessantes Ergebnis: In der Hauptschule zeigt sich während des Tests ein positiver Interaktionseffekt zwischen Testalgorithmus

und Fähigkeitsselbstkonzept auf die Erfolgserwartung ($\gamma_{21} = 0.14$, $p < .05$; Tabelle 11.4), im Gymnasium hingegen ein negativer Interaktionseffekt ($\gamma_{21} = -0.10$, $p < .01$; Tabelle 11.6). Die Simple-Slope-Analysen zeigen, dass sich dieser Effekt in der Hauptschule ebenfalls auf die Personen mit hohem Fähigkeitsselbstkonzept zurückführen lässt. Sie sind im CAT signifikant höher motiviert als im FIT ($\gamma_{21} = 0.12$, $p < .05$; widerspricht Hypothese 6a). Bei niedrigem Fähigkeitsselbstkonzept zeigen die Hauptschülerinnen und Hauptschüler im CAT und im FIT eine vergleichbar niedrige Erfolgserwartung ($\gamma_{21} = -0.08$, $p = .18$; widerspricht Hypothese 6b). Im Gymnasium stellt sich der Zusammenhang umgekehrt dar: Die Personen mit hohem Fähigkeitsselbstkonzept sind im CAT ähnlich hoch motiviert wie im FIT ($\gamma_{21} = -0.06$, $p = .09$; widerspricht Hypothese 6a). Gymnasiastinnen und Gymnasiasten mit niedrigem Fähigkeitsselbstkonzept zeigen jedoch im CAT eine signifikant höhere Erfolgserwartung als im FIT ($\gamma_{21} = 0.08$, $p < .05$; bestätigt Hypothese 6b). Nach dem Test verschwinden beide Effekte. In der Realschule sind keinerlei differentielle Interaktionseffekte zu finden.

Insgesamt sind nur wenige differentielle Effekte des Testalgorithmus auf die Motivation zur Testbearbeitung in Abhängigkeit vom Fähigkeitsselbstkonzept festzustellen. Hypothese 6 kann in ihrer Allgemeingültigkeit nicht bestätigt werden.

## 7.5 Effekte der Testinstruktion auf die Motivation zur Testbearbeitung

Mit Hypothese 7 wird vermutet, dass sich Interaktionseffekte zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Motivation zur Testbearbeitung zeigen, wenn die Testpersonen nicht über die Besonderheiten des Testalgorithmus aufgeklärt werden. In Hypothese 8 wird behauptet, dass diese Interaktionseffekte bei Aufklärung über den Testalgorithmus nicht bestehen. Daher wird im Folgenden separat für die Testbedingungen mit intransparenter und mit transparenter Testinstruktion geprüft, ob signifikante Interaktionseffekte von Fähigkeitsselbstkonzept und Testalgorithmus auf die Motivation zur Testbearbeitung vorliegen. Das im vorherigen Abschnitt beschriebene Gleichungssystem liegt auch den folgenden Hypothesentests zugrunde, wird jedoch separat für die beiden Stufen der UV Testinstruktion modelliert. Wie im vorherigen Abschnitt werden im Text lediglich jene Koeffizienten genannt, die für die Prüfung der Hypothesen relevant sind. Zudem werden nur die Ergebnisse für die Gesamtstichprobe exemplarisch in Tabellenform im Text dargestellt. Die vollständigen Ergebnisse der Mehrebenen-Wachstumsmodelle sind dem Anhang zu entnehmen (Tabelle 11.7 bis Tabelle 11.18). Eine deskriptive Übersicht über die Ausprägung der Erfolgserwartung in den Testbedingungen mit intransparenter beziehungsweise transparenter Testinstruktion bietet Tabelle 7.13.

*Tabelle 7.13*: Mittelwerte und Standardabweichungen der Erfolgserwartung bei intransparenter und transparenter Testinstruktion.

| | | Während des Tests | | Nach dem Test | |
|---|---|---|---|---|---|
| Stichprobe | Testinstruktion | *M* | *SD* | *M* | *SD* |
| Gesamtstichprobe | | | | | |
| | Intransparent | 2.47 | 0.59 | 2.33 | 0.68 |
| | Transparent | 2.45 | 0.61 | 2.30 | 0.70 |
| Personen mit hoher Hoffnung auf Erfolg | | | | | |
| | Intransparent | 2.63 | 0.59 | 2.49 | 0.69 |
| | Transparent | 2.67 | 0.60 | 2.52 | 0.70 |
| Personen mit niedriger Hoffnung auf Erfolg | | | | | |
| | Intransparent | 2.14 | 0.68 | 1.99 | 0.74 |
| | Transparent | 2.02 | 0.56 | 1.90 | 0.65 |
| Hauptschülerinnen/-schüler | | | | | |
| | Intransparent | 2.45 | 0.63 | 2.35 | 0.72 |
| | Transparent | 2.40 | 0.64 | 2.29 | 0.72 |
| Realschülerinnen/-schüler | | | | | |
| | Intransparent | 2.45 | 0.61 | 2.29 | 0.70 |
| | Transparent | 2.41 | 0.56 | 2.20 | 0.66 |
| Gymnasiastinnen/ Gymnasiasten | | | | | |
| | Intransparent | 2.55 | 0.55 | 2.42 | 0.62 |
| | Transparent | 2.55 | 0.61 | 2.45 | 0.69 |

Ohne die Gesamtstichprobe nach Leistungsmotiv oder Schulart zu unterteilen, zeigen sich weder bei intransparenter (widerspricht Hypothese 7) noch bei transparenter Testinstruktion (bestätigt Hypothese 8) statistisch signifikante Interaktionseffekte (vgl. Tabelle 7.14 und Tabelle 7.15).

*Tabelle 7.14*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.02 | 106.84 | < .01 | 2.31 | 0.03 | 81.85 | < .01 |
| SK ($\gamma_{01}$) | 0.36 | 0.05 | 7.78 | < .01 | 0.30 | 0.05 | 5.77 | < .01 |
| H ($\gamma_{02}$) | 0.24 | 0.05 | 4.84 | < .01 | 0.27 | 0.07 | 4.15 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.02 | -3.07 | < .01 | -0.03 | 0.03 | -0.91 | .36 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.03 | 0.03 | 0.95 | .34 | 0.07 | 0.04 | 1.84 | .07 |
| SK ($\gamma_{21}$) | 0.00 | 0.03 | 0.01 | .99 | 0.01 | 0.03 | 0.17 | .86 |
| H ($\gamma_{22}$) | -0.06 | 0.05 | -1.26 | .21 | -0.03 | 0.07 | -0.47 | .64 |

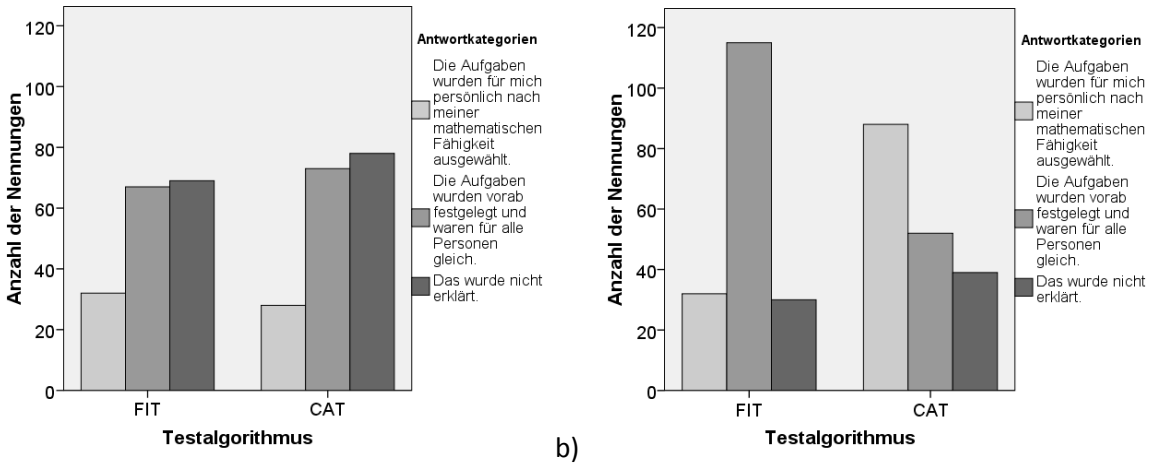
*Anmerkungen.* Während des Tests: *ICC* = .71; $R^2$ = .32; nach dem Test: *ICC* = .63; $R^2$ = .22;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 7.15*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.03 | 97.96 | < .01 | 2.37 | 0.03 | 74.63 | < .01 |
| SK ($\gamma_{01}$) | 0.45 | 0.04 | 11.81 | < .01 | 0.33 | 0.05 | 6.89 | < .01 |
| H ($\gamma_{02}$) | 0.26 | 0.04 | 6.23 | < .01 | 0.29 | 0.04 | 7.16 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.03 | -2.68 | < .01 | -0.10 | 0.04 | -2.55 | < .05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.02 | 0.02 | -1.01 | .31 | -0.04 | 0.03 | -1.18 | .24 |
| SK ($\gamma_{21}$) | -0.04 | 0.03 | -1.07 | .29 | 0.01 | 0.05 | 0.22 | .82 |
| H ($\gamma_{22}$) | -0.01 | 0.05 | -0.24 | .81 | -0.08 | 0.06 | -1.30 | .19 |

*Anmerkungen.* Während des Tests: *ICC* = .75; $R^2$ = .44; nach dem Test: *ICC* = .56; $R^2$ = .25;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

Angesichts dieser Ergebnisse stellt sich die Frage, ob die Testpersonen die Manipulation der Testinstruktion überhaupt rezipiert haben. Um diese Frage zu beantworten, werden in Abbildung 7.5 (1. Teststunde) und Abbildung 7.6 (2. Teststunde) die Antworthäufigkeiten auf die Frage zur Manipulationskontrolle dargestellt.



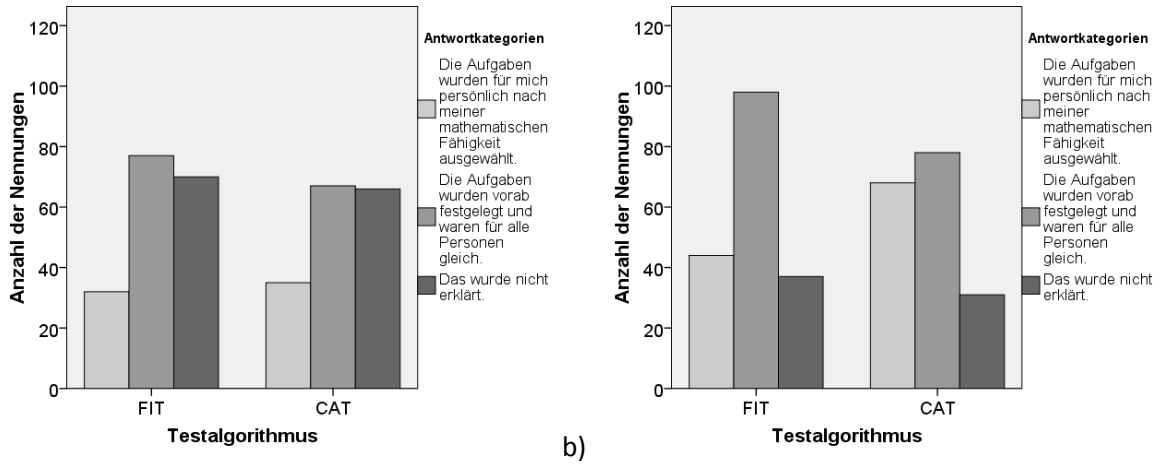

a) b)

*Abbildung 7.5*: Häufigkeitsverteilungen zur Frage, wie die Aufgaben des Mathematiktests in der ersten Teststunde ausgewählt wurden, bei intransparenter Testinstruktion (Abbildung 7.5a) und bei transparenter Testinstruktion (Abbildung 7.5b).

Wie in Abbildung 7.5 und Abbildung 7.6 zu sehen, fallen die Antworthäufigkeiten auf die Frage zur Auswahl der Mathematikaufgaben je nach Testinstruktion unterschiedlich aus. Bei intransparenter Testinstruktion gibt ein Großteil der Testpersonen (1. Teststunde: 42 %; 2. Teststunde: 39 %) richtigerweise an, dass nicht erklärt wurde, wie die Aufgaben ausgewählt werden. Ein ähnlich großer Anteil an Testpersonen (1. Teststunde: 40 %; 2. Teststunde: 41 %) vermutet allerdings, dass alle Testpersonen dieselben Aufgaben bearbeiten. Dies bestätigt die theoretische Annahme, dass die Testpersonen mit einer Erwartungshaltung in den Test gehen, die vermutlich auf früheren Erfahrungen mit (in der Regel nicht-adaptiven) Mathematiktests besteht. Unabhängig von dem tatsächlich hinterlegten Testalgorithmus nimmt lediglich ein kleiner Teil der Testpersonen an (1. Teststunde: 17 %; 2. Teststunde: 19 %), dass die Mathematikaufgaben in Abhängigkeit der individuellen Fähigkeit ausgewählt werden.

Ein anderes Bild zeigt sich, wenn die Testpersonen vor Beginn des Tests über den Testalgorithmus aufgeklärt werden: Ein eher geringer Teil der Testpersonen behauptet, das Vorgehen der Aufgabenauswahl sei nicht erklärt worden (1. Teststunde: 19 %; 2. Teststunde: 19 %). Bei diesen Testpersonen hat die Manipulation der Testinstruktion offenbar nicht gewirkt. Von den übrigen Testpersonen gibt die Mehrzahl den Testalgorithmus korrekt an (1. Teststunde gesamt: 71 %; FIT: 78 %, CAT: 63 %; 2. Teststunde gesamt: 58 %; FIT: 69 %, CAT: 47 %).

Offenbar wurde die unterschiedliche Formulierung der Testinstruktion von der Mehrheit der Testpersonen rezipiert. Allerdings scheint der adaptive Testalgorithmus insbesondere ohne expliziten Hinweis auf die individuell angepasste Aufgabenauswahl nur eingeschränkt wahrgenommen zu werden. Dies könnte erklären, weshalb in der Gesamtstichprobe der bei intransparenter Testinstruktion erwartete Interaktionseffekt zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung nicht besteht.




a) b)

*Abbildung 7.6*: Häufigkeitsverteilungen zur Frage, wie die Aufgaben des Mathematiktests in der zweiten Teststunde ausgewählt wurden, bei intransparenter Testinstruktion (Abbildung 7.6a) und bei transparenter Testinstruktion (Abbildung 7.6b).

Die separate Durchführung der Mehrebenen-Wachstumsmodelle für die Leistungsmotivgruppen erbringt folgende Ergebnisse: Personen mit hoher Hoffnung auf Erfolg zeigen generell in der

Transparenzbedingung nach einem CAT eine signifikant geringere Erfolgserwartung als nach einem FIT ($\gamma_{20}$ = -0.10, $p$ < .05). Bei Personen mit geringer Hoffnung auf Erfolg deutet sich generell eine motivierende Wirkung von CAT an, wenn der Testalgorithmus nicht erläutert wird (kleiner, knapp nicht signifikanter Effekt; während des Tests: $\gamma_{20}$ = 0.14, $p$ = .06; nach dem Test: $\gamma_{20}$ = 0.17, $p$ = .08). Signifikante Interaktionseffekte von Fähigkeitsselbstkonzept und Testalgorithmus auf die Erfolgserwartung zeigen sich in keiner der beiden Leistungsmotivgruppen, weder bei intransparenter noch bei transparenter Testinstruktion.

Eine genauere Analyse der Zusammenhänge in den einzelnen Schularten eröffnet, dass sich in der Intransparenzbedingung in der Realschule ein Haupteffekt des Testalgorithmus auf die Erfolgserwartung zugunsten des CAT finden lässt ($\gamma_{20}$ = 0.13, $p$ < .01). Klärt man die Realschülerinnen und -schüler über den Testalgorithmus auf, verschwindet dieser Effekt, und die Jugendlichen sind in beiden Testbedingungen gleich motiviert ($\gamma_{20}$ = -0.01, $p$ = .84). Ein ähnliches Ergebnis findet sich in der Hauptschule: Während des Tests sind die Jugendlichen bei intransparenter Testinstruktion im CAT motivierter als im FIT ($\gamma_{20}$ = 0.11, $p$ < .01). Bei Aufklärung über die Besonderheiten von CAT wird diese Differenz jedoch nivelliert und kehrt sich sogar fast ins Gegenteil um ($\gamma_{20}$ = -0.08, $p$ = .06). Interessant ist, dass die bereits beschriebenen Interaktionseffekte zwischen Fähigkeitsselbstkonzept und Testalgorithmus, die sich während des Tests in der Hauptschule und im Gymnasium finden (Abschnitt 7.3), ebenfalls nur bei Intransparenz bestehen (Hauptschule: $\gamma_{21}$ = 0.24, $p$ < .01; Gymnasium: $\gamma_{21}$ = -0.13, $p$ < .01; bestätigt Hypothese 7). Die Ergebnisse der Simple-Slope-Analysen entsprechen dabei inhaltlich dem Muster der Ergebnisse aus Abschnitt 7.3. So berichten Hauptschülerinnen und Hauptschüler mit hohem Selbstkonzept sowie Gymnasiastinnen und Gymnasiasten mit niedrigem Selbstkonzept eine signifikant höhere Erfolgserwartung im CAT als im FIT (Hauptschule, hohes SK: $\gamma_{20}$ = 0.29, $p$ < .01; Gymnasium, niedriges SK: $\gamma_{20}$ = 0.10, $p$ < .05). Für Hauptschülerinnen und Hauptschüler mit niedrigem Selbstkonzept besteht kein Unterschied der Erfolgserwartung im CAT und im FIT (Hauptschule, niedriges SK: $\gamma_{20}$ = -0.05, $p$ = .20). Bei Gymnasiastinnen und Gymnasiasten deutet sich bei hohem Selbstkonzept ein motivationsmindernder Effekt von CAT gegenüber FIT an, der jedoch knapp nicht signifikant ist (Gymnasium, hohes SK: $\gamma_{20}$ = -0.08, $p$ = .06). Klärt man die Schülerinnen und Schüler über den Testalgorithmus auf, verschwinden die Interaktionseffekte in beiden Schularten (Hauptschule: $\gamma_{21}$ = 0.02, $p$ = .81; Gymnasium: $\gamma_{21}$ = -0.06, $p$ = .29; Tabelle 11.16 und Tabelle 11.18; bestätigt Hypothese 8).

Zu den Hypothesen 7 und 8 ist zusammenzufassen, dass es für die Auswirkungen des Testalgorithmus und des Fähigkeitsselbstkonzepts auf die Erfolgserwartung in der Gesamtstichprobe wider Erwarten keinen Unterschied macht, ob die Testpersonen über den Testalgorithmus aufgeklärt werden oder nicht. Dies mag daran liegen, dass ein Großteil der Testpersonen davon ausgeht, einen FIT zu bearbeiten, sofern keine Erläuterung des Testalgorithmus erfolgt. Ohne transparente Testinstruktion scheint nur ein relativ geringer Teil der Testpersonen die individuell angepasste Aufgabenauswahl im CAT bewusst wahrzunehmen. Eine separate Betrachtung der Schularten bestätigt jedoch für die Hauptschule und das Gymnasium, dass Interaktionseffekte von Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung bestehen, wenn die Testpersonen nicht wissen, welcher Testalgorithmus dem Test zugrundeliegt. Eine transparente Testinstruktion führt zu einer Vermeidung der differentiellen Effekte auf die Erfolgserwartung, so dass sich die motivationalen Ausprägungen im FIT und im CAT angleichen. Hypothese 7 kann daher

nur für die Hauptschule und das Gymnasium bestätigt werden. Hypothese 8 kann generell bestätigt werden, da sich in der Transparenzbedingung weder in der Gesamtstichprobe noch in den untersuchten Teilstichproben Interaktionseffekte zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung finden lassen.

## 7.6 Effekte der Testschwierigkeit auf die Motivation zur Testbearbeitung

Die Hypothesen 9 und 10 beziehen sich auf einen Vergleich eines „typischen" CAT, der eine mittlere Lösungswahrscheinlichkeit von 50 Prozent hat, mit einem „einfachen" CAT, dessen Lösungswahrscheinlichkeit auf etwa 70 Prozent erhöht ist. Um diese hohe Lösungswahrscheinlichkeit zu erreichen, wurde die Aufgabenauswahl des adaptiven Algorithmus in dem einfachen CAT manipuliert (vgl. Abschnitt 6.2.1). Das Ziel war, dass die mittlere Aufgabenschwierigkeit des CAT nicht 0.0, sondern etwa -0.5 beträgt, was sich in einer individuellen Lösungswahrscheinlichkeit von etwa 70 Prozent statt etwa 50 Prozent niederschlagen sollte. Die Schwierigkeitsmanipulation funktionierte wie erwartet: So beträgt die mittlere Aufgabenschwierigkeit über alle Testpersonen im typischen CAT $M = 0.03$ ($SD = 0.47$), während sie im einfachen CAT bei $M = -0.48$ ($SD = 0.44$) liegt. Aufgrund der Verschiebung der Personenparameterverteilung (Abschnitt 7.1) resultieren diese Aufgabenschwierigkeiten jedoch nicht in den intendierten Lösungswahrscheinlichkeiten. Stattdessen ergeben sich mittlere Lösungswahrscheinlichkeiten von $RP = .41$ ($SD = .17$) im typischen CAT und von $RP = .54$ ($SD = .18$) im einfachen CAT. Diese Lösungswahrscheinlichkeiten unterscheiden sich zwar signifikant voneinander ($F(1,698) = 98.68$; $p < .01$; mittlerer Effekt mit $\eta^2 = .12$), doch liegen wider Erwarten statt einer einfachen und einer typischen Testbedingung eine typische und eine schwierige Testbedingung vor. Aufgrund dieser Verschiebung können die beiden Hypothesen 9 und 10 nicht überprüft werden. Die Effekte der Testschwierigkeit auf die Erfolgserwartung werden dennoch untersucht, allerdings haben diese Analysen nunmehr explorativen Charakter. In Tabelle 7.16 sind deskriptive Angaben zur Erfolgserwartung separat für die beiden Schwierigkeitsbedingungen dargestellt.

Da sich Hypothese 9 ausschließlich auf den CAT bezieht, wird der Effekt der Testschwierigkeit auf die Erfolgserwartung zunächst innerhalb der CAT-Testbedingung analysiert. Dazu wird folgendes Gleichungssystem des Mehrebenen-Wachstumsmodells spezifiziert:

Ebene 1: $y_{ij} = \beta_{0j} + \beta_{1j}(MZP) + r_{ij}$

Ebene 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(SK) + \gamma_{02}(TS) + \gamma_{03}(H) + u_{0j}$
$\beta_{1j} = \gamma_{10} + u_{1j}$

wobei

| | | |
|---|---|---|
| | *y* | Erfolgserwartung |
| | *i* | Indikator für Messzeitpunkt |
| | *j* | Indikator für Testperson |
| | MZP | Messzeitpunkt |
| | SK | Fähigkeitsselbstkonzept Mathematik |
| | TS | Testschwierigkeit (0: mittel, 1: hoch) |
| | H | Hoffnung auf Erfolg |

*Tabelle 7.16*: Mittelwerte und Standardabweichungen der Erfolgserwartung bei mittlerer und hoher Testschwierigkeit.

| | | Während des Tests | | Nach dem Test | |
|---|---|---|---|---|---|
| Stichprobe | Testschwierigkeit | *M* | *SD* | *M* | *SD* |
| Gesamtstichprobe | | | | | |
| | Mittel | 2.47 | 0.60 | 2.32 | 0.67 |
| | Hoch | 2.46 | 0.60 | 2.31 | 0.71 |
| Personen mit hoher Hoffnung auf Erfolg | | | | | |
| | Mittel | 2.65 | 0.61 | 2.49 | 0.71 |
| | Hoch | 2.64 | 0.58 | 2.52 | 0.68 |
| Personen mit niedriger Hoffnung auf Erfolg | | | | | |
| | Mittel | 2.10 | 0.62 | 1.98 | 0.66 |
| | Hoch | 2.07 | 0.64 | 1.92 | 0.74 |
| Hauptschülerinnen/-schüler | | | | | |
| | Mittel | 2.45 | 0.61 | 2.31 | 0.67 |
| | Hoch | 2.39 | 0.66 | 2.32 | 0.77 |
| Realschülerinnen/-schüler | | | | | |
| | Mittel | 2.46 | 0.58 | 2.29 | 0.64 |
| | Hoch | 2.40 | 0.59 | 2.19 | 0.72 |
| Gymnasiastinnen/ Gymnasiasten | | | | | |
| | Mittel | 2.52 | 0.59 | 2.40 | 0.66 |
| | Hoch | 2.58 | 0.57 | 2.46 | 0.66 |

Wie in den vorherigen Abschnitten, wird das Ergebnis für die Gesamtstichprobe exemplarisch in Tabellenform im Text dargestellt. Die vollständigen Ergebnisse der Mehrebenen-Wachstumsmodelle für die Analysen zum CAT sind dem Anhang zu entnehmen (Tabelle 11.19 bis Tabelle 11.24). Die Auswertung der Frage zur Manipulationskontrolle zeigt, dass die Testpersonen die beiden CAT-Testbedingungen, die sich objektiv sowohl im Hinblick auf die Differenz der mittleren Aufgabenschwierigkeiten als auch im Hinblick auf die mittleren Lösungswahrscheinlichkeiten deutlich voneinander unterscheiden, subjektiv nicht als unterschiedlich schwierig bewerten ($F(1,702) = 0.02$; $p = .88$, $\eta^2 = .00$). Auch im Hinblick auf die Erfolgserwartung gibt es weder bei Betrachtung der Gesamtstichprobe (vgl. Tabelle 7.17) noch bei separater Betrachtung der Personen mit hohem und geringem Leistungsmotiv (vgl. Tabelle 11.20 und Tabelle 11.21 im Anhang) einen signifikanten Unterschied zwischen den beiden CAT-Bedingungen. Eine separate Analyse der Schularten eröffnet in der Realschule einen signifikanten negativen Effekt der Schwierigkeit auf die Erfolgserwartung: Während sich dieser Effekt während des Tests allenfalls andeutet ($\gamma_{02} = -0.07$, $p = .22$), zeigen sich nach Testende deutliche Einbußen der Erfolgserwartung im schwierigen CAT im Vergleich zum typischen CAT ($\gamma_{02} = -0.13$, $p < .05$; Tabelle 11.23).

*Tabelle 7.17*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.02 | 109.11 | < .01 | 2.36 | 0.03 | 82.61 | < .01 |
| SK ($\gamma_{01}$) | 0.41 | 0.04 | 11.38 | < .01 | 0.32 | 0.04 | 7.56 | < .01 |
| TS ($\gamma_{02}$) | -0.01 | 0.04 | -0.16 | .88 | -0.03 | 0.04 | -0.62 | .53 |
| H ($\gamma_{03}$) | 0.23 | 0.03 | 9.04 | < .01 | 0.26 | 0.04 | 7.19 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.02 | -3.30 | < .01 | -0.06 | 0.03 | -2.26 | < .05 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .38; nach dem Test: *ICC* = .59; $R^2$ = .24;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit; H: Hoffnung auf Erfolg.

In Anlehnung an Hypothese 10 werden abschließend die beiden Testbedingungen „mittlere Testschwierigkeit“ und „hohe Testschwierigkeit“ explorativ separat analysiert. Das zugrunde liegende Gleichungssystem entspricht dem in Abschnitt 7.3. Die vollständigen Ergebnisse der Mehrebenen-Wachstumsmodelle für die Analysen zum CAT sind dem Anhang zu entnehmen (Tabelle 11.25 bis Tabelle 11.36).

Bei Betrachtung der Gesamtstichprobe zeigen sich in keiner der beiden Schwierigkeitsbedingungen Interaktionseffekte zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung. Allerdings berichten Personen mit geringer Hoffnung auf Erfolg bei mittlerer Schwierigkeit im CAT eine signifikant höhere Erfolgserwartung als im FIT ($\gamma_{20}$ = 0.21, $p$ < .01; Tabelle 11.28). Zum anderen tritt der in Abschnitt 7.3 berichtete Interaktionseffekt nur bei mittlerer Schwierigkeit in Erscheinung ($\gamma_{21}$ = -0.15, $p$ < .01; Tabelle 11.27): Personen mit hoher Hoffnung auf Erfolg, die ein hohes Fähigkeitsselbstkonzept haben, berichten nach Testende eine signifikant geringere Erfolgserwartung im CAT als im FIT ($\gamma_{20}$ = -0.11, $p$ < .05), während Personen mit hoher Hoffnung auf Erfolg, aber geringem Fähigkeitsselbstkonzept sich in beiden Testalgorithmen eine vergleichbare Erfolgserwartung bescheinigen ($\gamma_{20}$ = 0.11, $p$ = .15). In der schwierigen Testbedingung zeigt sich dieser Interaktionseffekt nicht ($\gamma_{21}$ = -0.03, $p$ = .56; Tabelle 11.29).

Die separate Analyse der Zusammenhänge nach Schulart zeigt bei mittlerer Schwierigkeit in der Hauptschule einen positiven Haupteffekt des Testalgorithmus auf die Erfolgserwartung nach dem Test ($\gamma_{20}$ = 0.10, $p$ < .05; Tabelle 11.31). Im typischen, mittelschwierigen CAT ist die Erfolgserwartung in der Hauptschule demnach signifikant höher als im FIT. In den Gymnasien besteht während eines schwierigen Tests dagegen ein negativer Haupteffekt des Testalgorithmus auf die Erfolgserwartung: Hier fällt die Erfolgserwartung im CAT signifikant geringer aus als im FIT ($\gamma_{20}$ = -0.04, $p$ < .05; Tabelle 11.36). Darüber hinaus ergeben sich bei mittlerer Schwierigkeit in der Hauptschule und im Gymnasium signifikante Interaktionseffekte zwischen Testalgorithmus und Fähigkeitsselbstkonzept während des Tests (Hauptschule: $\gamma_{21}$ = 0.21, $p$ < .05, vgl. Tabelle 11.31; Gymnasium: $\gamma_{21}$ = -0.16, $p$ < .05, vgl. Tabelle 11.33). Das Bild, das sich nach Durchführung der Simple-Slope-Analysen eröffnet, entspricht dabei dem Muster, das bereits bei intransparenter Testinstruktion festgestellt wurde:

Hauptschülerinnen und Hauptschüler mit hohem Fähigkeitsselbstkonzept sowie Gymnasiastinnen und Gymnasiasten mit niedrigem Fähigkeitsselbstkonzept sind im CAT signifikant motivierter als im FIT (Hauptschule, hohes SK: $\gamma_{20} = 0.29$, $p < .05$; Gymnasium, niedriges SK: $\gamma_{20} = 0.23$, $p < .05$). Für Hauptschülerinnen und Hauptschüler mit niedrigem Fähigkeitsselbstkonzept sowie Gymnasiastinnen und Gymnasiasten mit hohem Fähigkeitsselbstkonzept unterscheidet sich die Erfolgserwartung im FIT und im CAT nicht (Hauptschule, niedriges SK: $\gamma_{20} = 0.00$, $p = .95$; Gymnasium, hohes SK: $\gamma_{20} = 0.00$, $p = .98$). Dieses Interaktionsmuster scheint demnach in der Hauptschule und im Gymnasium unter typischen Testbedingungen (intransparente Testinstruktion, mittlere Testschwierigkeit) sehr stabil zu sein.

Zusammenfassend ist festzustellen, dass eine Prüfung der Hypothesen 9 und 10 anhand der vorliegenden Daten nicht möglich ist. Explorative Analysen des Effekts einer mittelschwierigen gegenüber einer schwierigen CAT-Bedingung ergeben numerisch geringere Erfolgserwartungen bei steigender Schwierigkeit. Nur in der Realschule lässt sich dieser negative Effekt der Testschwierigkeit auf die Erfolgserwartung statistisch gegenüber dem Zufall absichern. Zugleich findet sich unter „typischen“ Bedingungen, also bei mittlerer Schwierigkeit, in der Hauptschule ein motivierender Effekt von CAT gegenüber FIT. Zudem sind in der Hauptschule und im Gymnasium Interaktionseffekte zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung zu finden. Diese gehen auf eine höhere Erfolgserwartung im CAT als im FIT bei Personen aus der Hauptschule mit hohem Selbstkonzept und Personen aus dem Gymnasium mit niedrigem Selbstkonzept zurück. Bei hoher Schwierigkeit wirkt CAT im Vergleich zu FIT im Gymnasium demotivierend. Personen mit hoher Hoffnung auf Erfolg berichten in einem typischen CAT mit mittlerer Schwierigkeit eine geringere Erfolgserwartung als in einem FIT mittlerer Schwierigkeit, wenn sie über ein hohes Fähigkeitsselbstkonzept verfügen. Haben sie ein niedriges Fähigkeitsselbstkonzept, zeigen diese Personen im CAT eine vergleichbare Erfolgserwartung zur Testbearbeitung wie im FIT. Eine generell höhere Motivation (in Form der Erfolgserwartung) bei einer Lösungswahrscheinlichkeit von etwa 40 Prozent gegenüber einer Lösungswahrscheinlichkeit von etwa 50 Prozent, wie von Meyer et al. (1976) berichtet, kann anhand der vorliegenden Daten nicht bestätigt werden.

# 8 Diskussion

In diesem Abschnitt werden die Ergebnisse zunächst kurz zusammengefasst (Abschnitt 8.1) und dann vor dem theoretischen Hintergrund und unter Bezugnahme auf andere empirische Studien reflektiert und kritisch diskutiert (Abschnitt 8.2). Um die Bedeutsamkeit der Ergebnisse einordnen und ihre Nützlichkeit bewerten zu können, wird abschließend unter Rückgriff auf die in Abschnitt 5 formulierten theoretischen und praktischen Beweggründe der Arbeit auf die theoretische Relevanz (Abschnitt 8.3) und die praktische Relevanz (Abschnitt 8.4) der Befunde eingegangen.

## 8.1 Zusammenfassung der Ergebnisse

Als Leistungstests wurden in der vorliegenden Studie Tests zur mathematischen Kompetenz eingesetzt, die aus Bildungsstandards-Aufgaben für den Mittleren Schulabschluss bestanden. Die Skalierung der Daten aus diesen Kompetenztests und die Verankerung der Aufgabenschwierigkeiten im Rahmen der Kalibrierung an der Bildungsstandards-Stichprobe ist erfolgreich verlaufen. Lediglich die Schwierigkeit einer einzigen Aufgabe in dem Test der ersten Teststunde musste frei geschätzt werden. Ansonsten funktionieren die Aufgaben in der vorliegenden Stichprobe und in der Kalibrierungsstichprobe hinreichend ähnlich. Allerdings ergibt sich eine Verschiebung der Lösungswahrscheinlichkeiten. Dies führt dazu, dass die Hypothesen 9 und 10, die sich auf Effekte der Lösungswahrscheinlichkeiten auf die Motivation zur Testbearbeitung beziehen, nicht geprüft werden können. Die Reliabilitäten der Mathematiktests befinden sich in mäßiger Höhe, sind jedoch für die Zwecke der vorliegenden Arbeit ausreichend.

Die Hypothesen zur ersten Fragestellung können fast durchweg bestätigt werden: Das theoretisch postulierte, aus dem Erwartung-Wert-Modell der Leistungsmotivation von Eccles und Wigfield (2002) abgeleitete und auf eine Testsituation adaptierte Erwartung-Wert-Modell der Motivation zur Testbearbeitung entspricht den empirischen Daten (Hypothese 1). Auch eine genauere Betrachtung einzelner Pfade bestätigt die Annahmen weitgehend: Während des Tests sind die beiden Komponenten der aktuellen Motivation, Erwartung und Wert, positiv korreliert (Hypothese 2). Nach dem Test konnte der numerisch vorhandene positive Zusammenhang der beiden Komponenten jedoch wider Erwarten nicht gegen den Zufall abgesichert werden. Für die Vorhersage der aktuellen Testleistung erweist sich die Erwartungskomponente als deutlich wichtiger als die Wertkomponente (Hypothese 3). Dies unterstützt die Annahme, dass insbesondere die Erwartung für die Hypothesentests zur zweiten Fragestellung inhaltlich von Bedeutung ist. Schließlich wird bestätigt, dass sowohl die situationsspezifische als auch die dispositionelle Leistungsmotivation für das Verständnis der motivationalen Prozesse in einer Leistungstestsituation relevant sind (Hypothese 4). Die erste Fragestellung, ob sich die in Abschnitt 3.2.2.4 vorgestellte Spezifikation des Erwartung-Wert-Modells der Leistungsmotivation von Eccles und Wigfield (2002) zur Erklärung von Motivation zur Testbearbeitung in einer Leistungstestsituation eignet, wird zustimmend beantwortet.

Bevor die Hypothesen zur zweiten Fragestellung geprüft wurden, konnte sichergestellt werden, dass keine Carryover-Effekte in Abhängigkeit der Reihenfolge von CAT und FIT bestehen. Bei

der Prüfung der faktoriellen Invarianz zwischen den Datenstrukturen der beiden Messzeitpunkte (1. bzw. 2. Teststunde) konnte zudem für die Erwartungskomponente strikte faktorielle Invarianz bestätigt werden. Daher konnten die Analysen des komplexen Innersubjektdesigns für diese abhängige Variable ohne psychometrische Bedenken vorgenommen werden. Die Wertkomponente hingegen erfüllte den Mindestanspruch starker faktorieller Invarianz nicht und wurde deshalb von den Analysen zur zweiten Fragestellung ausgeschlossen.

Die Motivation zur Testbearbeitung (in Form der Erfolgserwartung) nimmt generell wie erwartet über die zwei Teststunden hinweg ab. Dies ist vermutlich als allgemeiner Ermüdungseffekt zu deuten, welcher aufgrund des counterbalancierten Versuchsplans für die Interpretation der Effekte der UV-Manipulationen unbedenklich ist. Hinsichtlich der Ergebnisse zu den Hypothesen der zweiten Fragestellung zeigt sich allerdings ein heterogenes Bild. Wie vermutet, fällt die Erfolgserwartung im CAT und im FIT sowohl bei Betrachtung der Gesamtstichprobe als auch bei separater Betrachtung der Leistungsmotivgruppen und der Schularten gleich hoch aus. Damit ist Hypothese 5 bestätigt. Die Annahme, dass die Wirkung des Testalgorithmus auf die Erfolgserwartung vom Fähigkeitsselbstkonzept abhängt (Hypothese 6), kann jedoch in Bezug auf die Gesamtstichprobe nicht bestätigt werden. Auch macht es diesbezüglich keinen Unterschied, wie die Testinstruktion formuliert wurde: Weder bei intransparenter noch bei transparenter Testinstruktion zeigt sich in der Gesamtstichprobe ein Interaktionseffekt zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung. Die Befunde der Gesamtstichprobe widersprechen somit Hypothese 7, in der der beschriebene Interaktionseffekt bei intransparenter Testinstruktion vermutet wurde. Sie bestätigen jedoch Hypothese 8, die aussagt, dass dieser Interaktionseffekt bei transparenter Testinstruktion nicht besteht. Dass Hypothese 7 nicht bestätigt werden kann, scheint sich darauf zurückführen zu lassen, dass die Testpersonen ohne explizite Erläuterung des Testalgorithmus implizit angenommen haben, dass sie einen FIT bearbeiten. Nur ein kleiner Teil der Testpersonen, die einen CAT bearbeiteten, nahm diesen ohne expliziten Hinweis bewusst als CAT wahr.

In Ergänzung zur Prüfung der Hypothesen an der Gesamtstichprobe wurden die Analysen explorativ separat für die Leistungsmotivgruppen „Hohe Hoffnung auf Erfolg“ und „Niedrige Hoffnung auf Erfolg“ sowie für die Schularten Hauptschule, Realschule und Gymnasium durchgeführt. Mit Blick auf die Leistungsmotivgruppen ergeben sich vereinzelt differentielle Effekte auf die Erfolgserwartung: So sind Personen mit hoher Hoffnung auf Erfolg in Abhängigkeit ihres mathematischen Fähigkeitsselbstkonzepts im CAT und im FIT unterschiedlich hoch motiviert. Dies lässt sich insbesondere darauf zurückführen, dass die Personen mit hoher Hoffnung auf Erfolg und hohem Fähigkeitsselbstkonzept im CAT eine signifikant geringere Erfolgserwartung berichten als im FIT. Besonders deutlich zeigt sich dies in einem typischen CAT mit 50-prozentiger Lösungswahrscheinlichkeit und in einem CAT mit transparenter Testinstruktion. Das heißt, Personen mit überdauerndem annäherndem Leistungsverhalten und hohem Fähigkeitsselbstkonzept werden durch den Hinweis, dass sie nach einer richtigen Antwort eine schwierigere Aufgabe erhalten und nach einer falschen Antwort eine einfachere Aufgabe und sich die Aufgabenauswahl somit an ihr Antwortverhalten anpasst, demotiviert. Bei Personen mit hoher Hoffnung auf Erfolg, aber niedrigem mathematischen Fähigkeitsselbstkonzept, die also trotz des allgemeinen leistungsannähernden Verhaltens speziell von ihrer mathematischen Kompetenz nur wenig überzeugt sind, ist die Erfolgserwartung im CAT und im FIT vergleichbar hoch. Personen mit gering ausgeprägter Hoffnung

auf Erfolg berichten im typischen CAT mit mittlerer Schwierigkeit eine signifikant höhere Erfolgserwartung als im FIT.

Eine separate Analyse der Schularten unterstützt die Vermutung, dass eine Aufklärung über den Testalgorithmus differentielle Effekte auf die Erfolgserwartung zwischen CAT und FIT vermeiden kann. Ohne eine vorherige Erläuterung des Testalgorithmus zeigen Jugendliche aus der Haupt- und der Realschule im CAT eine höhere Erfolgserwartung als im FIT. Zudem besteht in der Hauptschule und im Gymnasium jeweils ein Interaktionseffekt zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung. Dieser Effekt tritt unter typischen Testbedingungen besonders deutlich in Erscheinung (intransparente Testinstruktion, mittlere Testschwierigkeit): Die Hauptschülerinnen und -schüler, die von ihrer mathematischen Kompetenz überzeugt sind, sind im CAT motivierter als im FIT. Halten sie ihre mathematische Kompetenz jedoch für gering, fällt die Erfolgserwartung im CAT vergleichbar niedrig aus wie im FIT. Bei den Gymnasiastinnen und Gymnasiasten verhält es sich andersherum: Haben sie ein niedriges Fähigkeitsselbstkonzept, hat CAT gegenüber FIT eine motivierende Wirkung. Sind sie jedoch von ihrer mathematischen Kompetenz überzeugt, deutet sich numerisch eine motivationsmindernde Wirkung von CAT an. Eine Erläuterung des Testalgorithmus führt zu einem Ausgleich der Erfolgserwartung in allen Testbedingungen. Eine Anhebung der Testschwierigkeit wirkt sich in der Realschule negativ auf die Erfolgserwartung aus. Im Gymnasium wird bei hoher Schwierigkeit im CAT eine geringere Erfolgserwartung als im FIT berichtet.

Die Hypothesen zur zweiten Fragestellung können somit nur teilweise bestätigt werden: Während Hypothese 5 bestätigt wird, müssen die Hypothesen 6 und 7 in ihrer Allgemeingültigkeit verworfen werden. Hypothese 8 kann für die Gesamtstichprobe bestätigt werden. Die Hypothesen 9 und 10 können anhand der vorliegenden Daten nicht geprüft werden. Zur Beantwortung der zweiten Fragestellung, welches Bedingungsgefüge zwischen den Test- und Personenmerkmalen im Hinblick auf die Motivation zur Testbearbeitung besteht, wird festgestellt: Ein typischer CAT mit mittlerer Schwierigkeit kann im Vergleich zu FIT bei Personen mit hoher Hoffnung auf Erfolg zu Einbußen in der Erfolgserwartung führen, bei Personen mit niedriger Hoffnung auf Erfolg jedoch zu Steigerungen der Erfolgserwartung. Bei Schülerinnen und Schülern aus der Realschule wirkt sich CAT gegenüber FIT bei intransparenter Testinstruktion einheitlich positiv auf die Erfolgserwartung aus. In der Hauptschule und im Gymnasium hängt der Effekt des Testalgorithmus auf die Erfolgserwartung bei typischen Testbedingungen (mittlere Schwierigkeit, intransparente Testinstruktion) vom Fähigkeitsselbstkonzept ab. So steigert CAT im Vergleich zu FIT bei Personen mit hohem Fähigkeitsselbstkonzept in der Hauptschule und bei Personen mit niedrigem Fähigkeitsselbstkonzept im Gymnasium die Erfolgserwartung. Eine Erläuterung des Testalgorithmus vorab kann diese Unterschiede verhindern. Bei Personen mit hoher Hoffnung auf Erfolg kann die transparente Testinstruktion jedoch zu einer Minderung der Erfolgserwartung im CAT gegenüber FIT führen.

## 8.2 Allgemeine Diskussion der Ergebnisse

Die Ergebnisse liefern interessante und nützliche Anhaltspunkte für theoretische und praktische Implikationen. Im Folgenden werden zunächst die Befunde zu dem theoretischen Modell diskutiert, das die motivationalen Prozesse erklärt, die in einer Testsituation ablaufen (Abschnitt 8.2.1).

Anschließend wird der Frage nachgegangen, ob es notwendig ist, motivationale Variablen bei der Interpretation von Leistungstestergebnissen zu berücksichtigen (Abschnitt 8.2.2). Daraufhin werden die Ergebnisse zum Einfluss von Personen- und Testmerkmalen auf die Motivation zur Testbearbeitung in Form der Erfolgserwartung kritisch diskutiert (Abschnitt 8.2.3). Abschließend werden die Grenzen der Studie aufgezeigt (Abschnitt 8.2.4).

### 8.2.1 Motivationale Prozesse in einer Testsituation

Das Erwartung-Wert-Modell der Motivation zur Testbearbeitung eignet sich, die motivationalen Prozesse während einer aktuellen Leistungssituation abzubilden. Dies ist nicht selbstverständlich zu erwarten gewesen. Denn von den Autoren des zugrunde liegenden Modells (Eccles & Wigfield, 2002) ist zwar theoretisch postuliert worden, dass dieses Modell auch kurzfristige Motivationsänderungen abzubilden vermag (Eccles, 2005), doch wurde dies nach Erkenntnis der Autorin empirisch noch nicht gezeigt. Der Befund, dass sich das Modell zur Vorhersage der Motivation in einer Testsituation ohne (Aufgaben-)Wahlmöglichkeit eignet, weist zudem die kritische Behauptung von Beckmann und Heckhausen (2006) zurück, dass der Nutzen jeglicher Erwartung-Wert-Modelle auf die Prognose von Wahlen oder Entscheidungen beschränkt sei.

Allerdings ist die Wertkomponente der aktuellen Motivation zur Testbearbeitung bei einer Anwendung des Modells auf Low-Stakes-Testsituationen ohne Aufgabenwahlmöglichkeit von untergeordneter Bedeutung, da der Wert insbesondere Wahl- oder Entscheidungsverhalten prognostiziert. Es ist daher zu erwägen, bei Bezug auf solche Testsituationen noch spezifischer vom „Erwartung-Modell der Motivation zur Testbearbeitung“ zu sprechen. In High-Stakes-Testsituationen hingegen gewinnt die Wertkomponente für die aktuelle Motivation zur Testbearbeitung und die investierte Anstrengung an Bedeutung (z. B. Thelk et al., 2009). Auch in Testsituationen mit Wahlmöglichkeit sollte die Wertkomponente im Modell berücksichtigt werden. Ein Beispiel sind so genannte selbst-adaptive Tests, in denen die Testperson selbst entscheidet, welche Aufgabe sie als nächstes bearbeitet (z. B. Wise, Ponsoda & Olea, 2002). Diese erhöhte Selbstständigkeit seitens der Testpersonen in Form der Entscheidung, ob eine schwierigere oder eine einfachere Aufgabe vorgegeben werden soll, kann laut Georgiadou et al. (2006) motivierend wirken (vgl. auch Lunz et al., 1994 zur Anspruchsniveausetzung). Im Falle dieser Leistungstests sollte das vollständige Erwartung-Wert-Modell der Motivation zur Testbearbeitung zum Verständnis der motivationalen Prozesse genutzt werden.

Diese Ausführungen machen die eigenständige Bedeutsamkeit der beiden Komponenten Erwartung und Wert deutlich, die im Risikowahl-Modell (Atkinson, 1964) noch nicht erkannt wurde. Die Ergebnisse bestätigen zudem zumindest für den Zeitpunkt während des Tests, dass zwischen der Erwartung und dem Wert einer erfolgreichen Testbearbeitung eine positive Korrelation besteht (vgl. Wigfield et al., 1997). Für den Zeitpunkt nach dem Test ist dieser positive Zusammenhang lediglich numerisch erkennbar. Möglicherweise hängt das Verfehlen der statistischen Signifikanz zu diesem Messzeitpunkt mit der Operationalisierung der beiden Motivationskomponenten zusammen. So wird die Erwartungskomponente nach dem Test lediglich über zwei Fragen, die Wertkomponente über vier Fragen erfasst. Die Reliabilität dieser Skalen fällt niedriger aus als die Reliabilität der

entsprechenden, längeren Skalen zum Messzeitpunkt während des Tests, was den Nachweis einer statistisch signifikanten Korrelation (bei konstantem Stichprobenumfang) erschwert haben könnte.

Die positive Korrelation zwischen Erwartung und Wert schließt nicht aus, dass eine Veränderung der Erwartungskomponente einer Veränderung der Wertkomponente zeitlich vorausgeht (vgl. Eccles & Wigfield, 2002). Ein Hinweis hierfür könnte sein, dass sich der prozentuale Anteil der durch das Modell erklärten Varianz der Erwartungs- und der Wertkomponente mit zunehmender Testdauer beziehungsweise über die Messzeitpunkte hinweg annähert: Während die Varianzaufklärung der Erwartungskomponente zu Beginn deutlich höher ist als die der Wertkomponente, sinkt dieser Anteil im Verlauf der Zeit, während der Anteil erklärter Varianz der Wertkomponente steigt. Dies mag darauf hindeuten, dass sich die Erwartung einer erfolgreichen Testbearbeitung schneller, und zwar bereits während des ersten Tests, an die aktuelle Situation anpasst, während die Anpassung des Werts einer erfolgreichen Testbearbeitung zeitlich später folgt. Eine weitere Unterstützung dieser Vermutung mag in der fehlenden faktoriellen Invarianz der Wertkomponente zu finden sein, die sich möglicherweise psychologisch erklären lässt: Falls sich der Wert einer erfolgreichen Testbearbeitung träger verändert als die Erwartung einer erfolgreichen Testbearbeitung, mag es für die Testpersonen ermüdend und unnötig erscheinen, im Verlauf der zwei Teststunden viermal zu ihrer Wert-Einschätzung gefragt zu werden. Dies mag ein inkonsistentes Antwortverhalten hervorgerufen haben, das die faktorielle Invarianz der Wertkomponente verhindert hat (vgl. entsprechende Befunde zum Artefakt einer mangelhaften Reliabilität des TAT; Winter, John, Stewart, Klohnen & Duncan, 1998). Eine zeitlich nachgeordnete Anpassung der Wertkomponente, die sich an der aktuellen Erfolgserwartung orientiert, erscheint im Hinblick auf selbstwertdienliche Zwecke plausibel (vgl. Ausführungen in Abschnitt 3.2.2.2).

Die Befunde der vorliegenden Arbeit sprechen darüber hinaus dafür, dass sowohl das Fähigkeitsselbstkonzept als auch die Selbstwirksamkeitserwartung die Erwartungshaltung beeinflussen, mit der eine Person in eine Kompetenztestsituation geht. Die Feststellung von Eccles und Wigfield (1995), dass sich diese beiden Aspekte der Kompetenzüberzeugungen empirisch nicht trennen lassen, kann anhand der vorliegenden Daten nicht bestätigt werden. Beide Merkmale leisten einen eigenständigen Beitrag zur Vorhersage der Erwartung; Multikollinearität liegt zwischen den Merkmalen nicht vor (vgl. Abschnitt 7.2). Interessant wäre, ob sich, analog zum Interaktionseffekt zwischen Fähigkeitsselbstkonzept und Testalgorithmus auf die Erfolgserwartung, der für bestimmte Substichproben gefunden wurde, ein entsprechender Interaktionseffekt zwischen Selbstwirksamkeitserwartung und Testalgorithmus finden lässt. Möglicherweise zeigt sich eine solche Interaktion eher in der Gesamtstichprobe als in Substichproben, da sich die Selbstwirksamkeitserwartung im Gegensatz zum Fähigkeitsselbstkonzept unabhängig von sozialen oder dimensionalen Vergleichen entwickelt (Bong & Clark, 1999; Möller & Schiefele, 2004).

Die empirische Prüfung des Modells bestätigt außerdem das mehrfach theoretisch hervorgebrachte Postulat, dass für die Vorhersage der aktuellen Leistungsmotivation sowohl situationsspezifische als auch dispositionelle Aspekte beachtet werden sollten (z. B. Schmalt & Langens, 2009). Beide Aspekte der Leistungsmotivation sind für die Vorhersage der Testleistung bedeutsam, auch wenn sich der Einfluss des Leistungsmotivs lediglich indirekt äußert.

Das Erwartung-Wert-Modell der Motivation zur Testbearbeitung stellt auf dem Weg zu einem elaborierten Verständnis der motivationalen Prozesse in einer Testsituation einen nützlichen Schritt dar. Das Modell kann sich zu einer wichtigen theoretischen Grundlage entwickeln, motivationale

Prozesse in Schulleistungsstudien zu untersuchen und nachzuvollziehen. Dies erscheint insbesondere vor dem Hintergrund naheliegend, dass das Ausgangsmodell von Eccles und Wigfield (2002) speziell für den Einsatz im Schulbereich konzipiert wurde und das adaptierte Modell in der vorliegenden Studie an einer schulischen Stichprobe bestätigt wird. Zukünftig wäre es wünschenswert, die Gültigkeit des Erwartung-Wert-Modells der Motivation zur Testbearbeitung an weiteren Stichproben und unter Verwendung anderer Leistungstests, beispielsweise von High-Stakes-Tests oder Tests mit Aufgabenwahlmöglichkeit, zu überprüfen. In diesem Falle käme dem über die Wertkomponente laufenden Pfad des Modells eine große Bedeutung zu. Darüber hinaus sollte bei bestehender Wahlmöglichkeit auch die Motivkomponente Furcht vor Misserfolg in das Modell einbezogen werden, die in der vorliegenden Studie nicht berücksichtigt wurde (vgl. Ortner & Caspers, in press).

### 8.2.2 Bedeutsamkeit der Motivation zur Testbearbeitung für die Testleistung

Die Motivation zur Testbearbeitung erklärt in der vorliegenden Studie einen Anteil von 8 bis 12 Prozent der Testleistung. Dies entspricht dem Anteil der durch motivationale Variablen erklärten Leistungsvarianz, der in anderen empirischen Studien gefunden wurde (z. B. Robbins et al., 2004). Somit scheint der Einfluss der aktuellen Motivation auf die Leistung, beispielsweise unabhängig vom Inhalt des Leistungstests und von der Stichprobe, recht stabil zu sein. Auch wenn der Hauptanteil der Leistungsvarianz auf kognitive Variablen zurückzuführen ist, erscheint eine Beachtung der Motivation angesichts dieses Varianzanteils geboten (vgl. auch Steinmayr & Spinath, 2009). Denn Leistungstests werden in der Regel mit der Absicht durchgeführt, die maximale Leistung der Testpersonen zu erfassen (Cronbach, 1970). Daher lassen sich die erzielten Testergebnisse nur dann berechtigterweise als maximale Leistung interpretieren, wenn die Testpersonen hinreichend motiviert sind, ihre maximale Leistung zu zeigen (Thelk et al., 2009). Auch aus ethisch-psychologischer Sicht ist anzustreben, dass die Testpersonen motiviert sind und die Testsituation als angenehm empfinden (Häcker, Leutner & Amelang, 1998; Rost, 2000).

Die empirischen Befunde sprechen somit gegen kritische Stimmen, die eine Berücksichtigung motivationaler Variablen für die Leistungsvorhersage für vernachlässigbar halten (z. B. Gagné & St Père, 2001; Schmidt-Atzert, 2006; Spangler, 1992). Unterschiede in der Leistungsmotivation scheinen im Rahmen interindividueller Leistungsprognosen allerdings eine geringere Relevanz zu haben als im Rahmen intraindividueller Leistungsvorhersagen (Brunstein & Heckhausen, 2006). Dennoch ist eine Berücksichtigung motivationaler Einflussgrößen auf die Leistung angesichts der berichteten Ergebnisse auch in interindividuellen Vergleichen angeraten, insbesondere wenn es sich um Low-Stakes-Testsituationen handelt (z. B. Sundre & Kitsantas, 2004; vgl. Abschnitt 4.1). Angesichts der zahlreichen groß angelegten Vergleichsstudien unter Low-Stakes-Testbedingungen wie der internationalen PISA-Studie oder den nationalen Bildungsstandards-Erhebungen in Deutschland sollte der Motivation zur Testbearbeitung ausreichend Aufmerksamkeit geschenkt werden, um eine valide Interpretation der in diesen Studien gewonnenen Daten zu gewährleisten. Eine angemessene Berücksichtigung der Motivation kann beispielsweise erfolgen, indem die Befunde der vorliegenden Arbeit bei der Entwicklung eines Leistungstests für eine bestimmte Personengruppe beziehungsweise vor dem Einsatz eines Leistungstests beachtet werden, um äußere Merkmale des Tests wie die Testinstruktion für die jeweilige Stichprobe optimal motivierend zu gestalten (vgl. praktische

Empfehlungen in Abschnitt 8.4). Auf diese Weise kann erreicht werden, dass die Testergebnisse aus Tests, die die maximale Leistung von Personen erfassen sollen, auch in diesem Sinne valide interpretiert werden können und kein Konglomerat aus Leistung und Motivation darstellen, das sich einer validen Interpretation entzieht. Sind die Testergebnisse bereits erhoben, stellt das Aufspüren inkonsistenten Antwortverhaltens über Person-Fit-Analysen eine Möglichkeit dar, unmotivierte Testpersonen zu identifizieren (z. B. Meijer & Sijtsma, 2001; vgl. auch van Barneveld, 2007).

### 8.2.3 Einflussfaktoren auf die Motivation zur Testbearbeitung

Bei Betrachtung der Gesamtstichprobe ergibt sich kein Haupteffekt des Testalgorithmus auf die Motivation zur Testbearbeitung in Form der Erfolgserwartung. Dies war theoretisch angenommen worden. Es widerspricht jedoch auf den ersten Blick Befunden aus den wenigen anderen empirischen Untersuchungen zu diesem Thema, die von einem positiven Haupteffekt der Adaptivität auf die Motivation (Betz & Weiss, 1975, 1976a, 1976b) beziehungsweise von einem negativen Haupteffekt der Adaptivität auf die Motivation berichten (Frey et al., 2009; Ortner & Caspers, in press). Bei näherer Betrachtung gibt es jedoch Unterschiede zwischen den Studien, die die scheinbar divergierenden Ergebnisse erklären könnten.

So beschränken sich die genannten Studien auf die Untersuchung hoch leistungsfähiger Testpersonen (Studierende, Gymnasiastinnen und Gymnasiasten), während der vorliegenden Studie eine leistungsheterogene Stichprobe mit Jugendlichen aus den Schularten Hauptschule, Realschule, Integrierte Gesamtschule und Gymnasium zugrunde liegt. Wenn die Wirkung des Testalgorithmus auf die Erfolgserwartung auch in der vorliegenden Studie separat für die Schularten vorgenommen wird, zeigt sich (in der typischen Testbedingung ohne transparente Testinstruktion) ein positiver Effekt von CAT im Vergleich zu FIT auf die Erfolgserwartung in der Haupt- und Realschule. Im Gymnasium hingegen unterscheidet sich die Erfolgserwartung zwischen CAT und FIT nur dann, wenn die mittlere Schwierigkeit des Tests hoch ist. In dem Fall wirkt sich CAT im Vergleich zu FIT demotivierend auf die Gymnasiastinnen und Gymnasiasten aus, was auch die anderen genannten Studien berichten.

Die bisherigen empirischen Studien sind darüber hinaus kaum miteinander vergleichbar, weil sie Motivation zur Testbearbeitung sehr unterschiedlich operationalisiert haben. So erfassen Betz und Weiss (1976) Motivation über vier Fragen zu Anstrengung, Herausforderung und Wert, allerdings ohne nähere theoretische Angaben zu den Fragen zu machen. Frey et al. (2009) sowie Ortner und Caspers (in press) hingegen benutzen den FAM (Rheinberg et al., 2001), der vor allem die Erwartungskomponente der Motivation erfasst. Vor dem Hintergrund der Vielgestaltigkeit des Merkmals Motivation zur Testbearbeitung im Allgemeinen und der daraus resultierenden, jedoch unbeachteten Einschränkung der Vergleichbarkeit der bestehenden Studien im Speziellen tritt die Notwendigkeit zutage, im Rahmen einer wissenschaftlichen Auseinandersetzung mit Motivation zur Testbearbeitung stets zu verdeutlichen, auf welche Komponente der Motivation man sich bezieht.

Ein weiterer Unterschied zwischen den Studien, der die scheinbar widersprüchlichen Ergebnisse bedingt haben könnte, besteht in der Art des eingesetzten Leistungstests. Bei Frey et al. (2009) und bei Ortner und Caspers (in press) wurden rein kognitive, inhaltsferne Leistungstests eingesetzt, die die Konzentrationsleistung beziehungsweise das schlussfolgernde Denken erfassen. Bei Betz und Weiss (1975, 1976a, 1976b) wurde ein Vokabeltest eingesetzt, der Gedächtnisleistung

abfragt. In der vorliegenden Studie hingegen wurden domänenspezifische Kompetenztests in Form von Mathematiktests verwendet. Angesichts der Domänen- beziehungsweise Aufgabenspezifität von Kompetenzüberzeugungen wie dem Fähigkeitsselbstkonzept und der Selbstwirksamkeitserwartung sind in inhaltlichen Kompetenztests möglicherweise eher differentielle Effekte auf die Motivation zu erwarten als in den anderen Leistungstests. Diese differentiellen Effekte zeigen sich als Interaktionseffekte zwischen Kompetenzüberzeugungen und Testalgorithmus auf die Motivation. Im Hinblick darauf, dass viele Studien in der empirischen Bildungsforschung mit Kompetenztests arbeiten und Kompetenztests im Zentrum vieler groß angelegter Vergleichsstudien wie PISA stehen, sollte diesen möglichen Interaktionseffekten auf die Motivation Beachtung geschenkt werden.

Sowohl in der Hauptschule als auch im Gymnasium tritt unter typischen Testbedingungen (intransparente Testinstruktion, mittlere Testschwierigkeit) ein Interaktionseffekt zwischen Testalgorithmus und Fähigkeitsselbstkonzept auf die Erfolgserwartung auf. Dieser Effekt ist theoretisch erwartet worden, allerdings in Bezug auf die Gesamtstichprobe, also über die Schularten hinweg, formuliert worden. Angesichts der Tatsache, dass sich das Fähigkeitsselbstkonzept innerhalb des Bezugsrahmens der Schule entwickelt (*big fish little pond effect*, Marsh, 1993; vgl. auch Möller et al., 2009), erscheint es jedoch plausibel, dass sich mögliche Interaktionseffekte auf Schulartebene zeigen. Im Gymnasium ergibt sich der Interaktionseffekt in postulierter Form. So deutet sich hier eine geringere Erfolgserwartung im CAT als im FIT an, wenn die Jugendlichen über ein hohes Fähigkeitsselbstkonzept verfügen. Sind sie hingegen von ihren eigenen mathematischen Fähigkeiten nicht überzeugt, zeigen sie im CAT eine höhere Erfolgserwartung als im FIT. Dies lässt sich theoretisch folgendermaßen begründen: Es ist davon auszugehen, dass Personen mit einem hoch ausgeprägten Fähigkeitsselbstkonzept und einer hohen Leistungsausprägung ihre Testleistung in einem CAT als schlechter wahrnehmen als sie erwartet hatten. Denn die Testerfahrungen und damit die Erwartungshaltung, mit der diese Personen in die Testsituation gehen, beruhen in der Regel auf Tests mit nicht-adaptiven Testalgorithmen (Lunz & Bergstrom, 1994). Als empirisches Indiz für eine solche implizite Erwartungshaltung können die Ergebnisse zur Manipulationskontrolle interpretiert werden: Ohne eine Vorab-Erläuterung des Testalgorithmus vermuten die Testpersonen mehrheitlich, einen FIT zu bearbeiten. Die Testpersonen wissen nicht, dass die Aufgabenauswahl in einem CAT auf Basis des gezeigten Antwortverhaltens so erfolgt, dass jede Testperson im Mittel lediglich 50 Prozent der Aufgaben löst. Zudem orientieren sich Testpersonen bei der Wahrnehmung ihrer eigenen aktuellen Testleistung an der subjektiv empfundenen Anzahl richtig beantworteter Aufgaben, ohne jedoch die Schwierigkeit dieser Aufgaben dabei zu berücksichtigen (vgl. Abschnitt 4.2.2). Aus diesen beiden Gründen ist zu vermuten, dass die subjektiv wahrgenommene Testleistung im CAT die erwartete Testleistung der Gymnasiastinnen und Gymnasiasten mit hohem Fähigkeitsselbstkonzept unterschreitet und dass dies der Grund dafür ist, dass die Erfolgserwartung dieser Personen im CAT numerisch geringer ist als im FIT. Bei geringem Fähigkeitsselbstkonzept entsteht möglicherweise eine positive Diskrepanz zwischen der wahrgenommenen Leistung und der Erwartungshaltung im CAT.

In der Hauptschule wird der Interaktionseffekt bei intransparenter Testinstruktion zwar auch signifikant, allerdings entspricht das Muster der Interaktion nicht den Annahmen. Im Vergleich zu dem Effekt im Gymnasium stellt es sich in der Hauptschule umgekehrt dar. So sind Hauptschülerinnen und Hauptschüler mit einem hohen Fähigkeitsselbstkonzept im CAT motivierter als im FIT, während solche mit einem geringen Fähigkeitsselbstkonzept im CAT ähnlich gering motiviert sind wie im FIT. Möglicherweise lässt sich dieser Befund auf folgende Umstände

zurückführen: Die Jugendlichen in der Hauptschule weisen in der vorliegenden Studie absolut betrachtet eine leicht unterdurchschnittliche Erfolgserwartung auf. Möglicherweise befindet sich hier ein größerer Anteil an Personen, die keine Ambitionen haben, Leistung zu zeigen. Diejenigen Hauptschülerinnen und Hauptschüler aber, die von ihrer Kompetenz überzeugt sind, könnten durch die adaptive Aufgabenauswahl positiv überrascht und auf diese Weise motiviert sein. Denn diese Jugendlichen können in einem CAT etwa die Hälfte aller Aufgaben korrekt lösen. Es ist zwar zu vermuten, dass diese Personen innerhalb ihrer gewohnten Bezugsgruppe, der Hauptschulklasse, in Klassenarbeiten ebenfalls gute Ergebnisse erzielen. Dennoch dürfte die Erwartungshaltung dieser Jugendlichen, mit der sie in einen allgemeinen, schulartübergreifenden Kompetenztest gehen, angesichts des ihnen bekannten niedrigen absoluten Kompetenzniveaus in der Hauptschule eher gering ausfallen. Generell (d. h. bei einem durchschnittlich hoch ausgeprägten Fähigkeitsselbstkonzept) ist die Erfolgserwartung in der Hauptschule im CAT höher als im FIT.

Der Vorteil zugunsten des CAT in der Haupt- und Realschule sowie die berichteten Interaktionseffekte in der Hauptschule und im Gymnasium treten nur dann auf, wenn die Jugendlichen vor Beginn des Tests *nicht* über die Besonderheiten des verwendeten Testalgorithmus aufgeklärt werden. Eine Aufklärung über den verwendeten Testalgorithmus führt dazu, dass sowohl die Haupteffekte als auch die Interaktionseffekte nivelliert werden und im FIT und im CAT, unabhängig vom Fähigkeitsselbstkonzept, eine vergleichbar hohe Erfolgserwartung besteht. Aus diesen Ergebnissen lassen sich praktische Empfehlungen ableiten, wann ein Einsatz von CAT sinnvoll ist und wann nicht. Auf diese Empfehlungen wird in Abschnitt 8.4 eingegangen.

Die explorative Analyse des übergeordneten, zeitlich stabilen Leistungsmotivs Hoffnung auf Erfolg erbrachte zwei Befunde, die diskutiert werden sollten. Personen mit hoher Hoffnung auf Erfolg äußern im CAT im Vergleich zum FIT eine geringere Erfolgserwartung, wenn ihnen die Besonderheiten des Testalgorithmus vorab nähergebracht werden. Personen mit geringer Hoffnung auf Erfolg hingegen zeigen im typischen CAT mit mittlerer Schwierigkeit eine höhere Erfolgserwartung als im FIT. Dieser Befund hängt möglicherweise mit der Operationalisierung des Leistungsmotivs als respondentes Maß zusammen. Denn die Erfassung des Leistungsmotivs über ein selbstattribuiertes (Fragebogen-)Maß erfordert eine bewusste Reflektion über die eigene Leistungsorientierung. Dadurch entsteht eine Nähe zum Fähigkeitsselbstkonzept, wie Brunstein und Schmitt (2003) empirisch gezeigt haben (zitiert nach Brunstein & Heckhausen, 2006, S. 155; vgl. auch Halisch & Heckhausen, 1989). Weinert und Helmke (1997) berichten von einem engen Bezug zwischen allgemeiner Leistungsorientierung und Fähigkeitsselbstkonzept in Mathematik bereits bei Grundschulkindern. Es besteht Konsens, dass es sich bei respondent und implizit erfassten Leistungsmotiven um zwei verschiedene Aspekte der dispositionellen Leistungsmotivation handelt, die beide ihre Berechtigung haben (Brunstein, 2003; McClelland et al., 1989; Spangler, 1992). In Anbetracht dieser Erkenntnisse überrascht der Haupteffekt des Testalgorithmus auf die Erfolgserwartung zugunsten des CAT bei Personen mit geringer Hoffnung auf Erfolg nicht. Ebenso erklärbar wäre ein negativer Effekt von CAT im Vergleich zu FIT auf die Erfolgserwartung bei Personen mit hoher Hoffnung auf Erfolg bei *intransparenter* Testinstruktion. Dass sich dieser negative Effekt bei Personen mit hoher Hoffnung auf Erfolg in der vorliegenden Studie allerdings bei transparenter Testinstruktion zeigt, lässt eine alternative Erklärung naheliegend erscheinen: Möglicherweise bewirkt gerade der explizite Hinweis, dass einer richtig beantworteten Aufgabe eine schwierigere Aufgabe folgt, bei Personen mit stark ausgeprägtem leistungsannäherndem Verhalten eine Minderung der Erfolgserwartung.

Angesichts der vorangehenden Überlegungen erscheint es interessant, das Leistungsmotiv statt über einen Fragebogen implizit zu erfassen. Allerdings gibt es mehrere Befunde und Argumente, weshalb es sinnvoll ist, im Rahmen von Leistungstests das selbstattribuierte und nicht das implizite Leistungsmotiv zu berücksichtigen: Das implizite Leistungsmotiv sagt eher langfristige Verhaltenstendenzen und den Anstrengungseinsetz in „entspannten" Situationen ohne Leistungserwartungen von außen vorher, während sich das selbstattribuierte Leistungsmotiv eher für Verhaltensprognosen in spezifischen Situationen eignet, die mit Zeit- oder Leistungsdruck verbunden sind (Brunstein, 2003). Außerdem wird das implizite Leistungsmotiv vor allem durch inhaltliche Aufgabenmerkmale aktiviert, die die intrinsische Motivation der Testperson anregen, während sich das selbstattribuierte Leistungsmotiv eher durch soziale Anreize oder äußere Aufgabenmerkmale wie eine leistungsorientierte Testinstruktion aktivieren lässt (Lang & Fries, 2006; McClelland et al., 1989). Nach McClelland et al. (1989) empfiehlt es sich daher, in konkreten Leistungstestsituationen, in denen das Leistungsmotivmaß und das Leistungsverhaltensmaß in kurzem zeitlichen Anstand zueinander folgen und in denen die Testpersonen Aussagen über ihr Leistungsverhalten machen sollen, das selbstattribuierte Leistungsmotiv zu erheben.

Im Hinblick auf die mittlere Schwierigkeit des Tests konnten die formulierten Hypothesen nicht überprüft werden. Dennoch liefert auch die realisierte Schwierigkeitsmanipulation interessante Ergebnisse: So ist die Erfolgserwartung der Realschülerinnen und Realschüler in einem schwierigen CAT signifikant geringer als in einem CAT mit mittlerer Schwierigkeit. Dass sich dieser Befund lediglich in der Realschule zeigt, mag darauf zurückzuführen sein, dass die Kompetenz der meisten Jugendlichen dieser Schulart in mittlerer Höhe liegt. Da der CAT mit einer Aufgabe mittlerer Schwierigkeit beginnt und sich dann sukzessive an das Kompetenzniveau der Testpersonen anpasst, mag die Passung von Aufgabenschwierigkeit und Personenfähigkeit in der Realschule schneller erreicht worden sein als in den anderen Schularten. Der CAT in der vorliegenden Studie war mit zehn Aufgaben relativ kurz. Es kann nicht ausgeschlossen werden, dass sich ein Effekt der Schwierigkeit auf die Motivation deswegen am ehesten in der Realschule zeigt und sich dieser Effekt bei längeren Tests auch in den anderen Schularten finden lässt.

Zusammenfassend ist zur zweiten Fragestellung zu sagen, dass die differenzierte Betrachtung der Einflussfaktoren auf die Erfolgserwartung zwar ein komplexes, nicht einfach auszuwertendes Design mit sich gebracht hat, aber auch zu Erkenntnissen geführt hat, die bei einer oberflächlichen Betrachtung verborgen geblieben wären. Wie vermutet gibt es Anzeichen dafür, dass sich der Einsatz von CAT in Abhängigkeit von bestimmten Personen- und Testmerkmalen unterschiedlich auf die Erfolgserwartung auswirkt. Diese differentiellen Wirkungszusammenhänge könnten eine Erklärung für die inkonsistenten Ergebnisse bisheriger empirischer Studien zum Effekt von CAT auf die Motivation zur Testbearbeitung liefern (Betz & Weiss, 1976; Frey et al., 2009; Ortner & Caspers, in press). Ein Einsatz von CAT kann somit nicht generell empfohlen oder verworfen werden, sondern sollte im Einzelfall sorgfältig abgewogen werden (vgl. Abschnitt 8.4).

### 8.2.4 Methodische und inhaltliche Grenzen der Studie

Die vorliegende Arbeit hat weiterführende Ergebnisse hervorgebracht, sie unterliegt jedoch auch einigen methodischen und inhaltlichen Einschränkungen. So sind die eingesetzten Mathematiktests mit einer Länge von jeweils zehn Aufgaben im Vergleich zu den meisten Leistungstests als kurz zu bewerten. Es zeichnen sich zwar bereits bei dieser Länge Effekte der untersuchten Personen- und Testmerkmale auf die Erfolgserwartung ab. Es ist jedoch zu vermuten, dass sich diese Effekte bei längeren Tests stärker manifestieren. Eine Ausweitung der Testlänge war im Rahmen der vorliegenden Studie nicht möglich, da den Schulen eine Durchführungszeit von über zwei Schulstunden nicht zumutbar gewesen wäre. Für die Zukunft wären ähnliche Untersuchungen unter Verwendung längerer Leistungstests jedoch interessant, um abgesicherte Aussagen unabhängig von der Testlänge formulieren zu können.

Eine weitere Einschränkung ergibt sich durch die unzureichend geglückte Manipulation des Testmerkmals Schwierigkeit. Unerwarteterweise entsprechen die Lösungswahrscheinlichkeiten in der vorliegenden Stichprobe nicht den Lösungswahrscheinlichkeiten in der Kalibrierungsstichprobe, obwohl die Verteilung der Schülerinnen und Schüler auf die Schularten in beiden Stichproben ähnlich ist. Dieser Befund lässt sich vermutlich auf die unterschiedlichen Erhebungszeitpunkte zurückführen: Während die Datenerhebung für die Bildungsstandards-Stichprobe im Frühjahr und damit zu Beginn des zweiten Schulhalbjahres der neunten Klasse stattfand, geschah sie in der vorliegenden Studie zu Beginn des ersten Schulhalbjahres der neunten Klasse. Die Jugendlichen wurden also etwa sieben Monate früher getestet als die Jugendlichen der Vergleichsstichprobe. Es überrascht nicht, dass die verwendeten Mathematikaufgaben zu Beginn der neunten Klasse schwieriger sind als in der Mitte der neunten Klasse. Zur Verschiebung der Lösungswahrscheinlichkeiten mag darüber hinaus beigetragen haben, dass an der vorliegenden Studie lediglich Schülerinnen und Schüler aus Schleswig-Holstein teilgenommen haben, während die Bildungsstandards-Stichprobe aus Jugendlichen aller Bundesländer bestand. Bei PISA fällt die mittlere mathematische Kompetenz der Jugendlichen in Schleswig-Holstein etwas geringer aus als die mittlere mathematische Kompetenz in Deutschland (z. B. Frey, Asseburg, Ehmke & Blum, 2008; Neubrand et al., 2005). Da die Bildungsstandards-Aufgaben den PISA-Aufgaben ähnlich sind (Frey, Hartig & Carstensen, 2009), ist zu vermuten, dass sich entsprechende Unterschiede in der mittleren mathematischen Kompetenz auch auf die Ergebnisse eines Tests mit Bildungsstandards-Aufgaben auswirken. Aufgrund der Verschiebung der Lösungswahrscheinlichkeiten können die Daten der vorliegenden Studie nicht zur Prüfung der Hypothesen 9 und 10 verwendet werden, die sich auf Tests mit hoher Lösungswahrscheinlichkeit beziehen. Die explorativen Analysen haben zwar einige neue Erkenntnisse erbracht, doch steht eine Analyse der motivationalen Prozesse in einem „einfachen“ Test noch aus.

Die Operationalisierung der aktuellen Motivation zur Testbearbeitung über den OMQ (Boekaerts, 2002) wurde gewählt, weil der OMQ für den Einsatz während und unmittelbar nach einem Test konzipiert wurde. Einige Skalen dieses Instruments bestehen lediglich aus zwei Fragen, was nach Ansicht einiger Autoren für eine latente Modellierung zu wenig ist (z. B. Marsh, Byrne & Yeung, 1999). Diese Kritik wird jedoch mit Verweis auf Crombach et al. (2003) zurückgewiesen: Auch wenn eine Operationalisierung durch drei oder mehr Fragen pro Merkmal aus methodischer Sicht wünschenswert ist, ist eine latente Modellierung mit zwei Indikatoren möglich (vgl. auch Raykov & Marcoulides, 2006). Eine Verlängerung der Skalen des OMQ zöge den Nachteil mit sich, dass der aktuelle Testverlauf stärker unterbrochen würde. Somit erfordert der Anspruch, Motivation im

aktuellen Testverlauf zu messen, einen Kompromiss zwischen methodischem Anspruch und Praktikabilität des Einsatzes, der im OMQ in Form von teilweise sehr kurzen Skalen eingegangen worden ist.

Nicht optimal ist die Verwendung unterschiedlicher Instrumente zur Erfassung der Wertkomponente während des Tests und nach dem Test. Nach Erkenntnis der Autorin gibt es bislang kein bewährtes Instrument, das eine Erfassung des Werts über Fragen ermöglicht, die sich in ihrer Formulierung inhaltlich und sprachlich zu beiden Zeitpunkten entsprechen. Hier wäre eine Weiterentwicklung der bestehenden Instrumente zu begrüßen. Die Einschränkungen, die sich durch diesen Umstand für die vorliegende Arbeit ergeben, sind jedoch nicht grundlegend. Denn zum einen kommt der Wertkomponente in den meisten der hier untersuchten Hypothesen keine zentrale Rolle zu. Zum anderen werden die beiden Messzeitpunkte (während des Tests und nach dem Test) nicht direkt miteinander verglichen, so dass die unterschiedliche Operationalisierung methodisch unbedenklich ist.

Zu denken gibt allerdings die fehlende faktorielle Invarianz der Skalen, mit denen die Wertkomponente in der vorliegenden Arbeit erfasst wurde. Die Invarianz konnte für die Wertkomponente zu keinem der beiden Messzeitpunkte, weder während des Tests noch nach dem Test, bestätigt werden. Dies bedeutet, dass die Wertskalen in der ersten und der zweiten Teststunde unterschiedlich funktionieren, also voraussichtlich etwas Unterschiedliches messen und einen unterschiedlichen Validitätsbereich aufweisen könnten. Das Fehlen der Invarianz könnte durch psychologische Effekte der Wiederholungsmessungen bedingt sein, die in inkonsistentes Antwortverhalten aller oder einiger Testpersonen gemündet haben könnten (vgl. Ausführungen in Abschnitt 8.2.1). So nennt auch Lutz (2006) Unterschiede im Antwortstil zwischen den zu vergleichenden Gruppen (hier: Messzeitpunkte), die beispielsweise auf Motivationsdifferenzen oder Unterschiede in der Anstrengungsbereitschaft zurückgehen können, als eine typische Ursache für fehlende faktorielle Invarianz. Angesichts dieser Befunde ist zu hinterfragen, ob das Erwartung-Wert-Modell der Motivation zur Testbearbeitung in der ersten und der zweiten Teststunde das Gleiche abbildet. Eine Überprüfung des Modells anhand einer weiteren Stichprobe erscheint empfehlenswert. Zudem wäre es interessant, die Vermutung empirisch zu prüfen, dass die fehlende faktorielle Invarianz der Wertkomponente auf eine von den Testpersonen wahrgenommene Unangemessenheit der Wiederholungsmessung zurückzuführen ist.

Inhaltlich ergeben sich Grenzen der Studie, da innerhalb des experimentellen Settings nur solche Variablen und ihre Zusammenhänge erfasst werden konnten, die in einer Testsituation unmittelbar auf die Motivation zur Testbearbeitung und die Testleistung einwirken. Das Erwartung-Wert-Modell von Eccles und Wigfield (2002) ist jedoch wesentlich komplexer als das hier verwendete Modell. Somit könnte es interessant sein, auch indirekte Einflussgrößen auf die Motivation zur Testbearbeitung mit zu untersuchen, die beispielsweise zur Entwicklung des Fähigkeitsselbstkonzepts oder des Leistungsmotivs beitragen. Diese komplexe Sichtweise hätte jedoch den Rahmen der Möglichkeiten dieser Arbeit überstiegen und erschien zur Beantwortung der Fragestellungen nicht notwendig.

Weitere inhaltliche Aspekte, deren Berücksichtigung für das Thema der vorliegenden Arbeit vielversprechend gewesen wäre, die aber aus organisatorischen Gründen außer Acht gelassen werden mussten, sind emotionale Merkmale und Attributionen. Bezüglich der emotionalen

Merkmale wird im Zusammenhang mit Motivation zur Testbearbeitung beispielsweise häufig die Testängstlichkeit behandelt (z. B. Ortner & Caspers, in press; Tonidandel et al., 2002). Attributionen, also subjektive Ursachenzuschreibungen für die eigene Testleistung, scheinen in einem Leistungstest ohne Wahlmöglichkeit für die aktuelle Motivation sehr bedeutsam zu sein und nehmen möglicherweise eine wichtigere Rolle ein als das Leistungsmotiv (vgl. auch Möller & Schiefele, 2004). Um diese subjektiven Interpretationen besser zu verstehen und gegebenenfalls Ansätze für Interventionen entwickeln zu können, wäre es hochinteressant, in zukünftigen Studien auch die leistungsbezogenen Attributionen zu erfassen.

Die berichteten Ergebnisse müssen in Anbetracht der hier aufgeführten Grenzen interpretiert und bewertet werden. Trotz der methodischen und inhaltlichen Einschränkungen leistet die Arbeit einen wichtigen Beitrag zum besseren Verständnis motivationaler Prozesse in einer Leistungstestsituation und zu Auswirkungen verschiedener Personen- und Testmerkmale auf die Motivation zur Testbearbeitung. Der wesentliche theoretische und praktische Nutzen der Arbeit wird in den folgenden beiden Abschnitten abschließend zusammenfassend dargestellt.

## 8.3 Theoretische Relevanz der Ergebnisse

Der theoretische Beweggrund dieser Arbeit war, dass bislang kein theoretisches, wissenschaftlich fundiertes Modell vorliegt, das die motivationalen Prozesse erklärt, die in einer Testsituation ablaufen. Ein differenziertes Verständnis dieser Prozesse erscheint nützlich, da die Motivation zur Testbearbeitung durchschnittlich etwa zehn Prozent der Leistungsvarianz in einem Test erklärt. Für eine berechtigte Interpretation der Testergebnisse als maximale Leistung der Testpersonen ist es demzufolge notwendig, dass die Testpersonen während der Testbearbeitung hinreichend motiviert sind.

In der vorliegenden Arbeit wurde ein Ausschnitt des bewährten Erwartung-Wert-Modells der Leistungsmotivation (Eccles & Wigfield, 2002) als Grundlage genommen, um ein auf eine Testsituation spezifiziertes Erwartung-Wert-Modell der Motivation zur Testbearbeitung abzuleiten. Dieses spezifische Modell konnte in der vorliegenden Arbeit empirisch bestätigt werden. Damit liegt ein wichtiger Baustein für eine umfassende psychologische Theorie der Motivation zur Testbearbeitung vor, die ein grundlegendes Verständnis der motivationalen Prozesse in einer Leistungstestsituation ermöglicht. Das vorgestellte Modell ist als nützliche theoretische Grundlage für andere Studien zur Motivation in Testsituationen zu bewerten.

## 8.4 Praktische Relevanz der Ergebnisse

Neben dem theoretischen Beweggrund beruht die Arbeit auch auf einem praktischen Beweggrund. Dieser entstand vor dem Hintergrund, dass sich die psychometrischen Vorteile von CAT insbesondere bei groß angelegten Studien wie PISA auszahlen würden und ein Einsatz von CAT im Rahmen solcher Studien diskutiert wird (OECD, 2006). Die motivationalen Effekte von CAT im Vergleich zu FIT wurden jedoch noch nicht hinreichend verstanden. Dies ist aber nötig, um valide Testwertinterpretationen

sicherzustellen. Daher wurden verschiedene mögliche Einflussfaktoren auf die Motivation zur Testbearbeitung experimentell manipuliert und unter Verwendung einer der PISA-Stichprobe ähnlichen Stichprobe von Neuntklässlerinnen und Neuntklässlern untersucht. Das Ziel war, eine wissenschaftlich fundierte Empfehlung geben zu können, wie ein Leistungstest gestaltet sein sollte, der zugleich eine hohe Messeffizienz und eine hohe Motivation zur Testbearbeitung ermöglicht.

Als Fazit lassen sich aus den Ergebnissen folgende Empfehlungen ableiten: Grundsätzlich sollte der Einsatz von Tests mit mittlerer Schwierigkeit dem Einsatz von Tests mit hoher Schwierigkeit vorgezogen werden. Besteht die Stichprobe, an der der Test durchgeführt werden soll, aus Jugendlichen aller Schularten, sollte der Testalgorithmus vor Beginn des Tests erläutert werden. Das bedeutet, die Testinstruktion sollte je nach Testalgorithmus darüber aufklären,

- dass alle Testpersonen unabhängig von ihrer Leistungsausprägung dieselben Aufgaben erhalten und dass dies bedeuten kann, dass einige Aufgaben für die Testperson viel zu einfach oder viel zu schwierig sind (FIT), beziehungsweise
- dass die Aufgaben für jede Testperson passend zu ihrer individuellen Leistungsausprägung ausgewählt werden, dass nach einer richtigen/falschen Antwort eine schwierigere/leichtere Aufgabe folgt und dass so vermieden wird, dass die Testperson Aufgaben bearbeiten muss, die für sie persönlich viel zu einfach oder viel zu schwierig sind (CAT).

Wird keine Erläuterung des Testalgorithmus vorgenommen, ist in einem FIT damit zu rechnen, dass Gymnasiastinnen und Gymnasiasten mit einem geringen Fähigkeitsselbstkonzept sowie Hauptschülerinnen und Hauptschüler mit einem hohen Fähigkeitsselbstkonzept nicht optimal motiviert sind und nicht ihre maximale Leistung zeigen. In einem CAT ohne transparente Testinstruktion ist zu befürchten, dass Gymnasiastinnen und Gymnasiasten mit einem hohen Fähigkeitsselbstkonzept demotiviert sind und unter ihrer maximalen Leistung bleiben. Eine transparente Testinstruktion löst die Ungleichheiten auf und führt dazu, dass keine Person aufgrund ihrer subjektiven Kompetenzüberzeugung durch den verwendeten Testalgorithmus benachteiligt wird. Nur dann kann von einer fairen Testung und von einer validen Interpretation der Testergebnisse als maximale Leistung ausgegangen werden. Besteht eine transparente Testinstruktion, sind CAT und FIT psychologisch gleichermaßen für den Einsatz an Stichproben aus unterschiedlichen Schularten geeignet. Angesichts der Messeffizienz-Vorteile von CAT würde sich ein Einsatz von CAT mit transparenter Testinstruktion in solchen Fällen empfehlen. Dies gilt auch für groß angelegte Vergleichsstudien wie PISA. Allerdings ist dies natürlich nur dann durchführbar, wenn für die Testung Computer zur Verfügung stehen.

Soll der Test jedoch lediglich an einer Schulart durchgeführt werden, hängt die Empfehlung zur optimalen Testkonfiguration von der Schulart ab. Im Gymnasium gilt dieselbe Empfehlung wie für die heterogene Stichprobe: Die Testpersonen sollten vorab über die Besonderheiten des Tests aufgeklärt werden. Ist dies der Fall, können CAT und FIT gleichermaßen zu fairen und validen Testergebnissen führen. In der Realschule hingegen sollte der Test nach Möglichkeit in Form eines CAT durchgeführt werden, allerdings *ohne* vorherige Erläuterung des Testalgorithmus. Denn dann profitieren in dieser Schulart alle Jugendlichen gleichermaßen vom CAT und zeigen eine höhere Erfolgserwartung als im FIT. Zwar sind auch in einem intransparenten FIT keine Hinweise auf differentielle Effekte und damit

auf eine Beeinträchtigung der Testfairness zu finden, aber es scheint, dass die Realschülerinnen und Realschüler im FIT nicht ihre maximale Leistung zeigen. Dies würde die Validität der Testergebnisse eines FIT einschränken, sofern diese als maximale Leistung interpretiert werden. Für die Hauptschule kann aufgrund der vorliegenden Daten keine eindeutige Empfehlung gegeben werden. Grundsätzlich zeigen alle Jugendlichen dieser Schulart im CAT ohne transparente Testinstruktion eine höhere Erfolgserwartung als im FIT. Dies spricht für einen Einsatz eines intransparenten CAT in der Hauptschule und lässt vermuten, dass Hauptschülerinnen und Hauptschüler in einem intransparenten FIT nicht ihre maximale Leistung zeigen. Allerdings nimmt man mit einem intransparenten CAT an der Hauptschule in Kauf, dass Personen mit einem geringen Fähigkeitsselbstkonzept von dem Testalgorithmus weniger stark profitieren als Personen mit einem hohen Fähigkeitsselbstkonzept. Hier bedarf es weiterer Untersuchungen, um die uneinheitlichen Ergebnisse zu prüfen und besser zu verstehen.

Erschwert wird die Entscheidung für oder gegen einen Einsatz von CAT durch Effekte auf die aktuelle Motivation, die sich in Abhängigkeit vom dispositionellen Leistungsmotiv ergeben (vgl. Abschnitt 8.2.3). Da das dispositionelle Leistungsmotiv jedoch vor dem Einsatz eines Leistungstests in der Regel nicht bekannt ist und jede Testkonfiguration mit Nachteilen für bestimmte Motivausprägungen verbunden ist, erscheint es nicht zielführend, dieses Personenmerkmal bei der Empfehlung zu berücksichtigen. Eine Orientierung an den Schularten, die untersucht werden sollen, erscheint dagegen sinnvoll und praktikabel. Die genannten Empfehlungen sollten bei der Planung einer Testung in die Entscheidung eines Einsatzes von CAT oder FIT und in die Konzeption der Testinstruktion einfließen.

# 9 Resümee und Ausblick

Die beiden Hauptanliegen der vorliegenden Arbeit, die in Form zweier Fragestellungen formuliert wurden (vgl. Abschnitt 5), wurden erfolgreich bearbeitet. Die erste Fragestellung, die die Gültigkeit des Erwartung-Wert-Modells der Motivation zur Testbearbeitung kritisch hinterfragte, kann zustimmend beantwortet werden. Mit dem Erwartung-Wert-Modell der Motivation zur Testbearbeitung liegt nun ein empirisch bestätigtes theoretisches Modell vor, das auf dem bewährten Erwartung-Wert-Modell der Leistungsmotivation von Eccles und Wigfield (2002) beruht und die motivationalen Prozesse in einer Leistungstestsituation angemessen abbildet (vgl. Abschnitt 8.3). Die zweite Fragestellung, welches Bedingungsgefüge zwischen Test- und Personenmerkmalen im Hinblick auf die Motivation zur Testbearbeitung besteht, hat zu der Erkenntnis geführt, dass Empfehlungen hinsichtlich einer optimalen Konfiguration eines Leistungstests in Abhängigkeit der zu untersuchenden Stichprobe unterschiedlich ausfallen sollten (vgl. Abschnitt 8.4).

Im Hinblick auf groß angelegte Vergleichsstudien, die mit leistungsheterogenen Stichproben arbeiten, stellt sich CAT nicht nur aus psychometrischer Sicht, sondern auch aus psychologisch-motivationaler Sicht als vielversprechend dar. Allerdings sollten die Testpersonen vorab über die Besonderheiten dieses Testalgorithmus aufgeklärt werden. Unter dieser Voraussetzung unterstützen die Befunde der vorliegenden Arbeit grundsätzlich die Ansicht, dass „CAT is a useful tool in testing performance and shows promise in becoming one of the basic testing procedures especially in large-scale examination for licensing and certification purposes" (Georgiadou et al., 2006, S. 276; vgl. auch Lunz et al, 1994; Zenisky & Sireci, 2002). Allerdings setzt ein Einsatz von CAT voraus, dass Computer am Testort für die Testung zur Verfügung stehen. Außerdem empfiehlt es sich, vor Einführung eines CAT beziehungsweise vor einer Umstellung eines FIT auf einen CAT die mit der Implementation beziehungsweise Umstellung verbundenen Kosten gegenüber dem zu erwartenden Nutzen sorgfältig abzuwägen (vgl. Frey & Seitz, in press; Zara, 1999).

Um den Prozess der subjektiven Interpretation der eigenen Testleistung während der Testbearbeitung gründlicher zu verstehen und gegebenenfalls Ansätze zu Interventionen entwickeln zu können, wäre es wünschenswert, zukünftig zusätzlich die Attributionen der eigenen Leistung zu untersuchen (z. B. Weiner, 1994). Außerdem sollten die motivationalen Effekte einer Anhebung der mittleren Lösungswahrscheinlichkeit im CAT von 50 Prozent auf etwa 70 Prozent analysiert werden. Diese Schwierigkeitsmanipulation ist ohne größere psychometrische Nachteile durchführbar und könnte die Motivation zur Testbearbeitung bei allen Testpersonen positiv beeinflussen.

# 10 Literaturverzeichnis

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172.

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks (CA): Sage Publications.

Andrich, D. (1995). Review of the book: Computerized adaptive testing. A primer. *Psychometrika, 60*, 615-620.

Arvey, R. D., Strickland, W., Drauden, G. & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695-716.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*, 359-372.

Atkinson, J. W. (1964). *An introduction to motivation*. New York (NY): Van Nostrand.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York (NY): W. H. Freeman and Company.

Battle, E. S. (1966). Motivational determinants of academic competence. *Journal of Personality and Social Psychology, 4,* 634-642.

Baumert, J., Artelt, C., Klieme, E. & Stanat, P. (2001). PISA: Programme for International Student Assessment. Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 285-310). Weinheim: Beltz.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441-462.

Beckmann, J. & Heckhausen, H. (2006). Motivation durch Erwartung und Anreiz. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 105-142). Heidelberg: Springer.

Beckmann, J. & Keller, J. A. (2009). Risikowahl-Modell. In V. Brandstätter & J. H. Otto (Hrsg.), *Handbuch der Allgemeinen Psychologie - Motivation und Emotion* (11. Aufl., S. 120-125). Göttingen: Hogrefe.

Bergstrom, B. A., Lunz, M. E. & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137-149.

Betz, N. E. (1975). Prospects: New types of information and psychological implications. In D. J. Weiss (Hrsg.), *Computerized adaptive trait measurement: Problems and prospects* (Research Report 75-5). Minneapolis (MN): University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E. & Weiss, D. J. (1976a). *Effects of immediate knowledge of results and adaptive testing on ability test performance* (Research Report 76-3). Minneapolis (MN): University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E. & Weiss, D. J. (1976b). *Psychological effects of immediate knowledge of results and adaptive ability testing* (Research Report 76-4). Minneapolis (MN): University of Minnesota, Department of Psychology, Psychometric Methods Program.

Binet, A. & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormeaux [Neue Methoden zur Diagnostik des Intelligenzniveaus bei Anormalen]. *L'Année Psychologique, 11*, 191-244.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Boekaerts, M. (2002). The On-Line Motivation Questionnaire: A self-report instrument to assess students' context sensitivity. *New Directions in Measures and Methods, 12*, 77-120.

Bong, M. & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist, 34*, 139-153.

Brandstätter, V. (2009). Persistenz und Zielablösung. In V. Brandstätter & J. H. Otto (Hrsg.), *Handbuch der Allgemeinen Psychologie: Motivation und Emotion* (S. 79-88). Göttingen: Hogrefe.

Braun, E., Woodley, A., Richardson, J. T. & Leidner, B. (2011). *Comparing questionnaires of self rated competences with research of questionnaire development*. Manuscript submitted for publication.

Breckler, S. J., Olson, J. M. & Wiggins, E. C. (2006). *Social psychology alive.* Belmont (CA): Thomson Wadsworth.

Brookhart, S. M., Walsh, J. M. & Zientarski, W. A. (2006). The dynamics of motivation and effort for classroom assessments in middle school science and social studies. *Applied Measurement in Education, 19*, 151-184.

Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist, 34*, 75-85.

Brown, S. M. & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*, 133-136.

Brunstein, J. C. (2003). Implizite Motive und motivationale Selbstbilder: Zwei Prädiktoren mit unterschiedlichen Gültigkeitsbereichen. In J. Stiensmeier-Pelster & F. Rheinberg (Hrsg.), *Diagnostik von Motivation und Selbstkonzept* (2. Aufl., S. 59-88). Göttingen: Hogrefe.

Brunstein, J. C. (2006). Implizite und explizite Motive. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 235-253). Heidelberg: Springer.

Brunstein, J. C. & Heckhausen, J. (2006). Leistungsmotivation. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 143-191). Heidelberg: Springer.

Brunstein, J. C. & Hoyer, S. (2002). Implizites versus explizites Leistungsstreben: Befunde zur Unabhängigkeit zweier Motivationssysteme. *Zeitschrift für Pädagogische Psychologie, 16*, 51-62.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte Auflage). München: Pearson Studium.

Cole, J. S., Bergin, D. A. & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624.

Crombach, M. J., Boekaerts, M. & Voeten, M. J. M. (2003). Online measurement of appraisals of students faced with curricular tasks. *Educational and Psychological Measurement, 63*, 96-111.

Cronbach, L. J. (1970). *Essentials of psychological testing*. New York: Harper & Row.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

Dahme, G., Jungnickel, D. & Rathje, H. (1993). Güteeigenschaften der Achievement Motives Scale (AMS) von Gjesme und Nygard (1970) in der deutschen Übersetzung von Göttert und Kuhl: Vergleich der Kennwerte norwegischer und deutscher Stichproben. *Diagnostica, 39*, 257-270.

De Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York (NY): Guilford Press.

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39*, 223-238.

Dweck, C. S., Mangels, J. A. & Good, C. (2004). Motivational effects on attention, cognition, and performance. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (S. 41-56). Mahwah (NJ): Lawrence Erlbaum Associates.

Eccles, J. S. (1993). School and family effects on the ontogeny of children's interests, self-perceptions, and activity choices. In J. E. Jacobs (Ed.), *Nebraska Symposium on Motivation, 1992: Developmental Perspectives on Motivation* (pp. 145-208). J. E. Jacobs. Lincoln (NE): University of Nebraska Press.

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105-121). New York: Guilford Press.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L. et al. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75-146). San Francisco (CA): W. H. Freeman.

Eccles, J. S. & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin, 21*, 215-115.

Eccles, J. S. & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.

Eccles, J. S., Wigfield, A., Harold, R. & Blumenfeld, P. B. (1993). Age and gender differences in children's self and task perceptions during elemantry school. *Child Development, 64*, 830-847.

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379-393.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden.* Weinheim: Beltz.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (Vol. 1). Hamburg: IEA-ETS Research Institute.

Frey, A. (2006). *Validitätssteigerungen durch adaptives Testen.* Frankfurt am Main: Peter Lang.

Frey, A. (2007). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 261-278). Heidelberg: Springer.

Frey, A., Asseburg, R., Ehmke, T. & Blum, W. (2008). Mathematische Kompetenz im Ländervergleich. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 127-147). Münster: Waxmann.

Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 169-184.

Frey, A., Hartig, J. & Carstensen, C. H. (2009). *Validierung des Tests zur Messung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss.* Manuskript eingereicht zur Publikation.

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica, 55*, 20-28.

Frey, A., Hartig, J. & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement, 28*, 39-53.

Frey, A. & Moosbrugger, H. (2004). Kann die Konfundierung von Konzentrationsleistung und Aktivierung durch adaptives Testen mit dem FAKT vermieden werden? *Zeitschrift für Differentielle und Diagnostische Psychologie, 25,* 1-17.

Frey, A. & Seitz, N.-N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. *Zeitschrift für Pädagogik, 56. Beiheft*, 40-51.

Frey, A., & Seitz, N.-N. (in press). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement*.

Frey, A., Seitz, N.-N. & Kröhne, U. (2010). *Reporting differentiated literacy results in PISA by the use of multidimensional adaptive testing.* Manuscript submitted for publication.

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J. et al. (2009). *PISA 2006 Skalenhandbuch. Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality, 40*, 21-34.

Gagné, F. & St Père, F. (2001). When IQ is controlled, does motivation still predict achievement? *Intelligence, 30*, 71-100.

Geiser, C. (2010). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung*. Wiesbaden: Verlag für Sozialwissenschaften.

Georgiadou, E., Triantafillou, E. & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology, 37*, 261-278.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L. & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Guay, F., Marsh, H. W. & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95,* 124-136.

Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen*. Bern: Hans Huber.

Halisch, F. & Heckhausen, H. (1989). Motive-dependent versus ability-dependent valence functions for success and failure. In F. Halisch & J. H. L. van den Bercken (Eds.), *International Perspectives on achievement and task motivation* (pp. 51-67). Amsterdam: Swets & Zeitlinger.

Hambleton, R. K. & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14*, 75-96.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park (CA): Sage Publications.

Harmon, D. J. (2010). *Multiple perspectives on computer adaptive testing for K-12 assessments*. Paper presented at the National Conference on Student Assessment. 17.11.2010 from http://ccsso.confex.com/ccsso/2010/webprogram/Session1359.html.

Hartig, J. & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*, 57-63.

Häusler, J. & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quaterly, 50*, 75-87.

Heckhausen, H. (1980). *Motivation und Handeln*. Berlin: Springer.

Heckhausen, J. & Heckhausen, H. (2006). Motivation und Handeln: Einführung und Überblick. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 1-9). Heidelberg: Springer.

Hox, J. J. (2000). Multilevel analyses of grouped and longitudinal data. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples* (pp. 15-32). Mahwah (NJ): Lawrence Erlbaum Associates.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jerusalem, M. (1984). *Selbstbezogene Kognitionen in schulischen Bezugsgruppen. Eine Längsschnittstudie. Band 1 des Berichts über das Forschungsvorhaben*. Berlin: Freie Universität Berlin.

Jopt, U.-J. (1978). *Selbstkonzept und Ursachenerklärung in der Schule. Zur Attribuierung von Schulleistungen.* Bochum: Kamp.

Kanfer, F. (1987). Selbstregulation und Verhalten. In H. Heckhausen, P. M. Gollwitzer & F. E. Weinert (Hrsg.), *Jenseits des Rubikon: Der Wille in den Humanwissenschaften* (S. 286-299). Berlin: Springer.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R. et al. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*, 139-151.

Keppel, G. & Wickens, T. D. (2004). *Design and analysis. A researcher's handbook*. Upper Saddle River, New Jearsey: Pearson Prentice Hall.

Kessels, U., Rau, M. & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology, 76,* 761-780.

Kim, J. & McLean, J. E. (1995). *The influence of examinee test-taking motivation in computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME). ERIC Document Reproduction Service No. ED392839.

Kingsbury, G. G., & Hauser, C. (2004). *Computerized adaptive testing and no child left behind*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego (CA).

Kingsbury, G. G. & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 93-115). New Jersey (NJ): Lawrence Erlbaum Associates.

Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptionelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III - Dritte internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit.* (S. 85-133). Opladen: Leske + Budrich.

Köller, O. (2004). *Konsequenzen von Leistungsgruppierungen*. Münster: Waxmann.

Köller, O. (2010). Bildungsstandards. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (3. Aufl., S. 529-550). Wiesbaden: Verlag für Sozialwissenschaften.

Köller, O., Daniels, Z., Schnabel, K. U. & Baumert, J. (2000). Kurswahlen von Mädchen und Jungen im Fach Mathematik: Zur Rolle von fachspezifischem Selbstkonzept und Interesse. *Zeitschrift für Pädagogische Psychologie, 14*, 26-37.

Köller, O. & Möller, J. (1995). Kontrafaktisches Denken nach schulischen Erfolgen und Mißerfolgen. *Zeitschrift für Pädagogische Psychologie, 9*, 105-110.

Krützer, B. & Probst, H. (2006). *IT-Ausstattung der allgemeinen bildenden und berufsbildenden Schulen in Deutschland. Bestandsaufnahme 2006 und Entwicklung 2001 bis 2006*. Berlin: Bundeministerium für Bildung und Forschung (BMBF).

Kubinger, K. D. (1995). *Einführung in die Psychologische Diagnostik*. Weinheim: Psychologie Verlags Union.

Kuhl, J. (1983). Leistungsmotivation: Neue Entwicklungen aus modelltheoretischer Sicht. In H. Thomae (Hrsg.), *Psychologie der Motive* (S. 505-625). Göttingen: Hogrefe.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M. et al. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente. Materialien aus der Bildungsforschung Nr. 72.* Berlin: Max-Planck-Institut für Bildungsforschung.

Lang, J. W. B. & Fries, S. (2006). A revised 10-item version of the Achievement Motives Scale: Psychometric properties in German-speaking samples. *European Journal of Psychological Assessment, 22*, 216-224.

Lilley, M., Barker, T. & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education, 43*, 109-123.

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. MESA Memorandum No. 69. In S. Chae, U. Kang, E. Jeon & J. M. Linacre (Hrsg.), *Development of computerized middle school achievement test [In Korean]*. Seoul: Komesa Press.

Little, T. D., Schnabel, K. U. & Baumert, J. (2000). *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. Mahwah (NJ): Lawrence Erlbaum Associates.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale (NJ): Lawrence Erlbaum Associates.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks (CA): Sage Publications.

Lunz, M. E. & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31*, 251-263.

Lunz, M. E., Bergstrom, B. A. & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research, 21,* 623-634.

Lutz, M. T. (2006). *Validität und faktorielle Invarianz einer neuropsychologischen Testbatterie zur Intelligenzprüfung bei Patienten mit Epilepsie* [Online-Dissertation]. Zugriff am 26.03.2011 unter http://bieson.ub.uni-bielefeld.de/volltexte/2006/966/.

Marsh, H. W. (1993). Academic self-concept: Theory, measurement, and research. In J. Suls (Ed.), *Psychological perspectives on the self: The self in social perspective* (pp. 59-98). Hillsdale (NJ): Psychology Press.

Marsh, H. W., Byrne, B. M. & Yeung, A. S. (1999). Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist, 34*, 155-167.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*, 397-416.

Martens, T., Goldhammer, F., Rölke, H., Scharaf, A. & Upsing, B. (2008). Technology Based Assessment – ein Gemeinschaftsprojekt der Arbeitseinheiten „Informationszentrum Bildung“ und „Bildungsqualität und Evaluation“. *DIPF informiert, 12*, 2-6. Zugriff am 26.03.2011 unter http://www.dipf.de/de/publikationen/dipf-informiert.

McClelland, D. C., Atkinson, J. W., Clark, R. A. & Lowell, E. L. (1953). *The achievement motive*. New York (NY): Appleton-Century-Crofts.

McClelland, D. C., Koestner, R. & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96*, 690-702.

Meece, J. L., Wigfield, A. & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*, 60-70.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement, 25*, 107-135.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Meyer, W.-U. (1973). Anstrengungsintention in Abhängigkeit von Begabungseinschätzung und Aufgabenschwierigkeit. *Archiv für Psychologie, 125*, 245-262.

Meyer, W.-U. (1984). *Das Konzept von der eigenen Begabung*. Bern: Huber.

Meyer, W.-U., Folkes, V. & Weiner, B. (1976). The perceived informational value and affective consequences of choice behavior and intermediate difficulty task selection. *Journal of Research in Personality, 10*, 410-423.

Middleton, J. A. & Toluk, Z. (1999). First steps in the development of an adaptive theory of motivation. *Educational Psychologist, 34*, 99-112.

Miller, N. E. (1944). Experimental studies of conflict. In J. M. Hunt (Ed.), *Personality and the behavior disorders: A handbook based on experimental and clinical research* (Vol. I, pp. 431-465). New York (NY): Ronald Press.

Mills, C. N. & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75-99). New York (NY): Kluwer Academic Publishers.

Möller, J. (2008). Lernmotivation. In A. Renkl (Hrsg.), *Lehrbuch Pädagogische Psychologie* (S. 263-298). Bern: Huber.

Möller, J. & Köller, O. (2000). Spontaneous and reactive attributions following academic achievement. *Social Psychology of Education, 4*, 76-86.

Möller, J. & Köller, O. (2001). Dimensional comparisons: An experimental approach to the Internal/External Frame Of Reference Model. *Journal of Educational Psychology, 93*, 826-835.

Möller, J., Pohlmann, B., Köller, O. & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129-1167.

Möller, J. & Schiefele, U. (2004). Motivationale Grundlagen der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz* (S. 101-124). Wiesbaden: Verlag für Sozialwissenschaften.

Möller, J. & Trautwein, U. (2009). Selbstkonzept. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 179-203). Heidelberg: Springer.

Moosbrugger, H. (2002). Item-Response-Theorie (IRT). In M. Amelang & W. Zielinski (Hrsg.), *Psychologische Diagnostik und Intervention* (S. 68-92). Heidelberg: Springer.

Moosbrugger, H. (2007a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 216-239, 250-259). Heidelberg: Springer.

Moosbrugger, H. (2007b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 99-112). Heidelberg: Springer.

Moosbrugger, H. & Heyden, M. (1997). *Frankfurter Adaptiver Konzentrationsleistungs-Test.* Bern: Huber.

Moreno, K. E. (1997). CAT-ASVAB operational test and evaluation. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 199-205). Washington, DC: American Psychological Association (APA).

Murphy, P. K. & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology, 25*, 3-53.

Murray, H. A. (1938). *Explorations in personality*. New York (NY): Oxford University Press.

Muthén, B., & Asparouhov, T. (2010). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp.15-40). New York (NY): Routledge Academic.

Muthén, L. K. & Muthén, B. (2009). *Mplus short courses. Berlin 20 – 21 July 2009* [Arbeitsmaterial].

Neubrand, M., Blum, W., Ehmke, T., Jordan, A., Senkbeil, M., Ulfig, F. et al. (2005). Mathematische Kompetenz im Ländervergleich. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?* (S. 51-84). Münster: Waxmann.

Oakland, T. & Harris, J. G. (2009). Impact of test-taking behaviors on full-scale IQ scores from the Wechsler Intelligence Scale for Children-IV Spanish Edition. *Journal of Psychoeducational Assessment, 27*, 366-373.

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris, Frankreich: Autor.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris, Frankreich: Autor.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, Frankreich: Autor.

OECD (2006). *The OECD Programme for International Student Assessment* [Electronic Version]. Retrieved 02.02.2011 from http://www.pisa.oecd.org/dataoecd/51/27/37474503.pdf.

OECD. (2007). *PISA 2006. Science competencies for tomorrow` s world: Vol. 1*. *Analysis*. Paris, Frankreich: Autor.

OECD. (2010). *The OECD Programme for the International Assessment of Adult Competencies (PIAAC)* [Electronic Version]. Retrieved 04.03.2011 from http://www.oecd.org/dataoecd/13/45/41690983.pdf.

Olson, L. (2003). Legal twists, digital turns: Computerized testing feels the impact of 'no child left behind'. *Education Week's Technology Counts, 22*, 11-14, 16.

O'Neil, H. F., Sugrue, B. & Baker, E. L. (1996). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment, 3*, 135-157.

Ortner, T. M. & Caspers, J. (in press). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Parshall, C. G., Harmes, J. C., Davey, T. & Pashley, P. J. (2010). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215-230). New York (NY): Springer.

Pawlik, K. (1997). Unterstützung psychologischer Eignungsdiagnostik durch den Computer. Bewertung neuester Entwicklungen im Psychologischen Dienst der Bundeswehr. In K. Puzicha (Hrsg.), *Arbeitsberichte Psychologischer Dienst der Bundeswehr: Neue Wege in der Personalpsychologie* (S. 145-192). Bonn: Bundesministerium der Verteidigung.

Pintrich, P. R. & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33-40.

Pintrich, P. R. & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Upper Saddle River (NJ): Pearson Education.

Pintrich, P. R., Smith, D. A. F., Garcia, T. & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies For Learning Questionnaire (MLSQ). *Educational and Psychological Measurement, 53*, 801-813.

Prenzel, M. & Blum, W. (2007). *Entwicklung eines Testverfahrens zur Überprüfung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss*. Kiel: IPN.

Prochaska, M. (1998). *Leistungsmotivation: Methoden, soziale Erwünschtheit und das Konstrukt; Ansatzpunkte zur Entwicklung eines neuen eignungsdiagnostischen Verfahrens*. Frankfurt am Main: Peter Lang.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D. et al. (2006). *PISA 2003 Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

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

Raykov, T. & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Mahwah (NJ): Lawrence Erlbaum Associates.

Reinecke, J. (2005). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: Oldenbourg.

Rheinberg, F. (2008). *Motivation*. Stuttgart: Kohlhammer.

Rheinberg, F., Vollmeyer, R. & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen. *Diagnostica, 47,* 57-66.

Richter, T. (2007). Wie analysiert man Interaktionen von metrischen und kategorialen Prädiktoren? Nicht mit Median-Splits! *Zeitschrift für Medienpsychologie, 19,* 116-125.

Rink, K. (1994). *Motivationale und volitionale Determinanten des Leistungshandelns*. Aachen: Shaker.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

Röll, A. (1994). *Die motivationale Orientierung in schulischen Leistungssituationen und ihr Effekt auf das Leistungsergebnis - eine Längsschnittuntersuchung*. Düsseldorf: Heinrich-Heine-Universität Düsseldorf.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau, 50*, 140-156.

Rost, J. (2000). Allgemeine Standards für die Evaluationsforschung. In W. Hager, J. L. Patry & H. Brezing (Hrsg.), *Evaluation psychologischer Interventionsmaßnahmen: Standards und Kriterien* (S. 129-140). Bern: Hans Huber.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Hans Huber.

Rudolph, U. (2009). *Motivationspsychologie kompakt*. Weinheim: Beltz.

Satorra, A. & Bentler, P. M. (1999). *A scaled difference chi-square test statistic for moment structure analysis* [Electronic Version]. Retrieved 04.03.2011 from http://citeseerx.ist.psu.edu.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23-74.

Schiefele, U. (1996). *Motivation und Lernen mit Texten*. Göttingen: Hogrefe.

Schiefele, U. (2009). Motivation. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 151-177). Heidelberg: Springer.

Schmalt, H.-D. (1996). Zur Kohärenz von Motivation und Kognition. In J. Kuhl & H. Heckhausen (Hrsg.), *Motivation, Volition und Handlung* (S. 241-273). Göttingen: Hogrefe.

Schmalt, H.-D. & Langens, T. A. (2009). *Motivation*. Stuttgart: Kohlhammer.

Schmidt-Atzert, L. (2006). Leistungsrelevante Rahmenbedingungen/Leistungsmotivation. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 223-241). Heidelberg: Springer.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

Schmitt, M., Baumert, A. & Hofmann, W. (2007). Person, Situation oder Interaktion? - Eine zeitlose Streitfrage. In R. Frankenberger, S. Frech & D. Grimm (Hrsg.), *Politische Psychologie und politische Bildung* (S. 58-74). Schwalbach: Wochenschau Verlag.

Schneider, K. & Schmalt, H.-D. (1994). *Motivation*. Stuttgart: Kohlhammer.

Schunk, D. H., Pintrich, P. R. & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications*. Upper Saddle River (NJ): Pearson Education.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, pp. 429-438). Oxford: Elsevier.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton (FL): Chapman & Hall/CRC.

Smith, L. F. & Smith, J. K. (2002). Relation of test-specific motivation and anxiety to test performance. *Psychological Reports, 91*, 1011-1021.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introducation to basic and advanced multilevel modeling*. London: Sage Publications.

Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112*, 140-154.

Stanat, P. (2008, Juni). *Daten der Bildungsforschung: Aktueller Stand und neue Entwicklungen.* Vortrag auf der 4. Konferenz für Sozial- und Wirtschaftsdaten, Wiesbaden.

Statistisches Bundesamt (2010). *Bildung und Kultur: Allgemeinbildende Schulen. Schuljahr 2007/08. Fachserie 11, Reihe 1*. Zugriff am 11.01.2011 unter http://www-ec.destatis.de.

Steinmayr, R. & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences, 19*, 80-90.

Steyer, R. & Partchev, I. (2000). *Latent state-trait theory in computerized adaptive testing*. Paper presented at the 42nd Annual Conference of the International Military Testing Association. Retrieved Access Date: 23.02.2011 from http://www.zpid.de/redact/link.php?link=1522.

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and performance?* Paper presented at the Annual Meeting of the American Educational Research Association (AERA). ERIC Document Reproduction Service No. ED432588.

Sundre, D. L. (2007). *The Student Opinion Scale (SOS): A measure of examinee motivation. Test Manual*. Harrisonbourg (VA): The Center for Assessment & Research Studies.

Sundre, D. L. & Finney, S. J. (2002). *Enhancing the validity and value of learning assessment: Furthering the development of a motivation scale*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans (LA).

Sundre, D. L. & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6-26.

Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*, 8-9.

Thelk, A. D., Sundre, D. L., Horst, S. J. & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education, 58*, 129-151.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101-133). Mahwah (NJ): Lawrence Erlbaum Associates.

Tonidandel, S., Quinones, M. A. & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*, 320-332.

Trope, Y. (1975). Seeking information about one's own ability as a determinant of choice among tasks. *Journal of Personality and Social Psychology, 32*, 1004-1013.

Trope, Y. & Brickman, P. (1975). Difficulty and diagnosticity as determinants of choice among tasks. *Journal of Personality and Social Psychology, 31*, 918-925.

Trotter, A. (2003). A question of direction: 'Adaptive' testing puts federal officials and experts at odds. *Education Week's Technology Counts, 22*, 17-20.

Valentine, J. C., DuBois, D. L. & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*, 111-133.

Van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31*,31-46.

Van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

Van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Journal of Psychology, 216*, 3-11.

Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern Item Response Theory*. New York: Springer.

Van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). New York (NY): Springer.

Volz-Sidiropoulou, E. (2004). Computerbasierte Psychodiagnostik. In H.-J. Fisseni (Hrsg.), *Lehrbuch der psychologischen Diagnostik* (S. 279-297). Göttingen: Hogrefe.

Vygotsky, L. S. (1978). *Mind and society*. Cambridge (MA): Harvard University Press.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1-21.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 1-21). Mahwah (NJ): Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (2000). Item Response Theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 61-99). Mahwah (NJ): Lawrence Erlbaum Associates.

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrika, 54,* 427-450.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158-177.

Way, D. (2010). *Some perspectives on CAT for K-12 assessments*. Paper presented at the National Conference on Student Assessment. 17.11.2010 from http://ccsso.confex.com/ccsso/2010/webprogram/Session1359.html.

Weiner, B. (1994). *Motivationspsychologie*. Weinheim: Beltz.

Weinert, F. E. & Helmke, A. (1997). *Entwicklung im Grundschulalter.* Weinheim: Beltz PVU.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*, 70-84.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review, 6*, 49-78.

Wigfield, A. & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12*, 265-310.

Wigfield, A. & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Wigfield, A. & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91-120). San Diego (CA): Academic Press.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C. et al. (1997). Changes in children's competence beliefs and subjekctive task values across the elementary school years: A three-year study. *Journal of Educational Psychology, 89*, 451-469.

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C. & Duncan, L. E. (1998). Traits and Motives: Toward an integration of two traditions in personality research. *Psychological Review, 105*, 230-250.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*, 152-166.

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., Ponsoda, V. & Olea, J. (2002). Self-adapted testing: An overview. *International Journal of Continuing Engineering Education and Life-Long Learning, 12,* 107-122.

Wolf, L. F. & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.

Wolf, L. F., Smith, J. K. & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalized Item Response modelling software*. Victoria: ACER Press.

Zara, A. R. (1999). Using computerized adaptive testing to evaluate nurse competence for licensure: Some history and forward look. *Advances in Health Sciences Education, 4*, 39-48.

Zenisky, A. L. & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.

# 11 Anhang

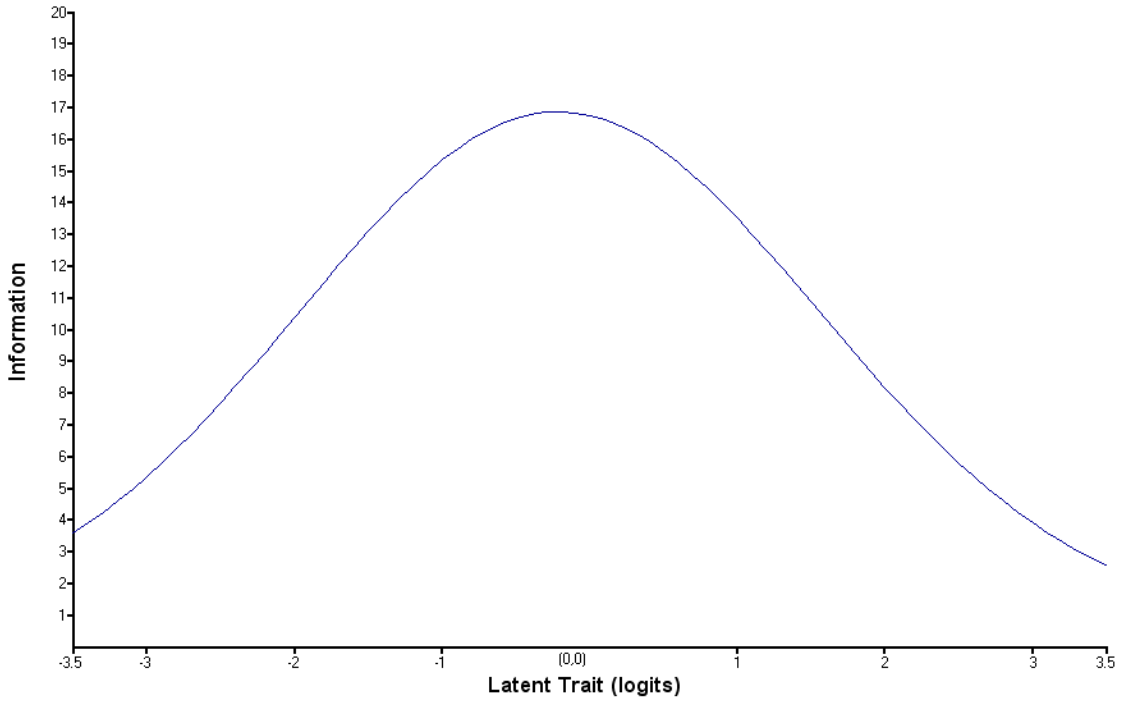



*Abbildung 11.1*: Testinformationsfunktion (*test information curve*, TIC) für die erste Teststunde.

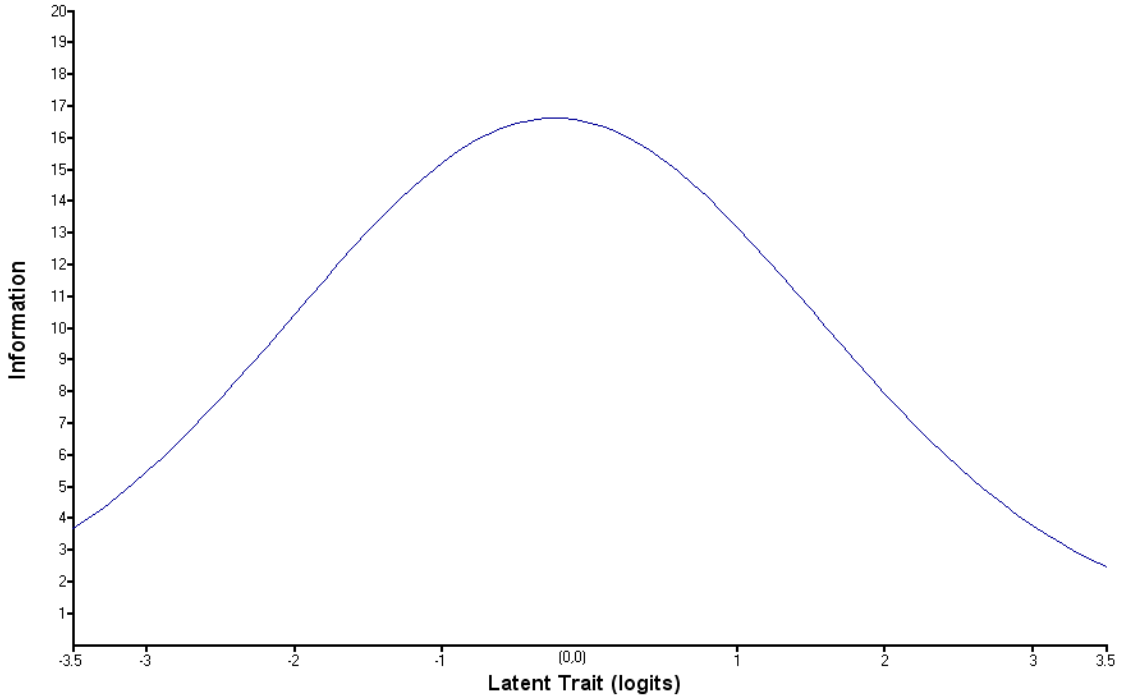



*Abbildung 11.2*: Testinformationsfunktion (*test information curve*, TIC) für die zweite Teststunde.

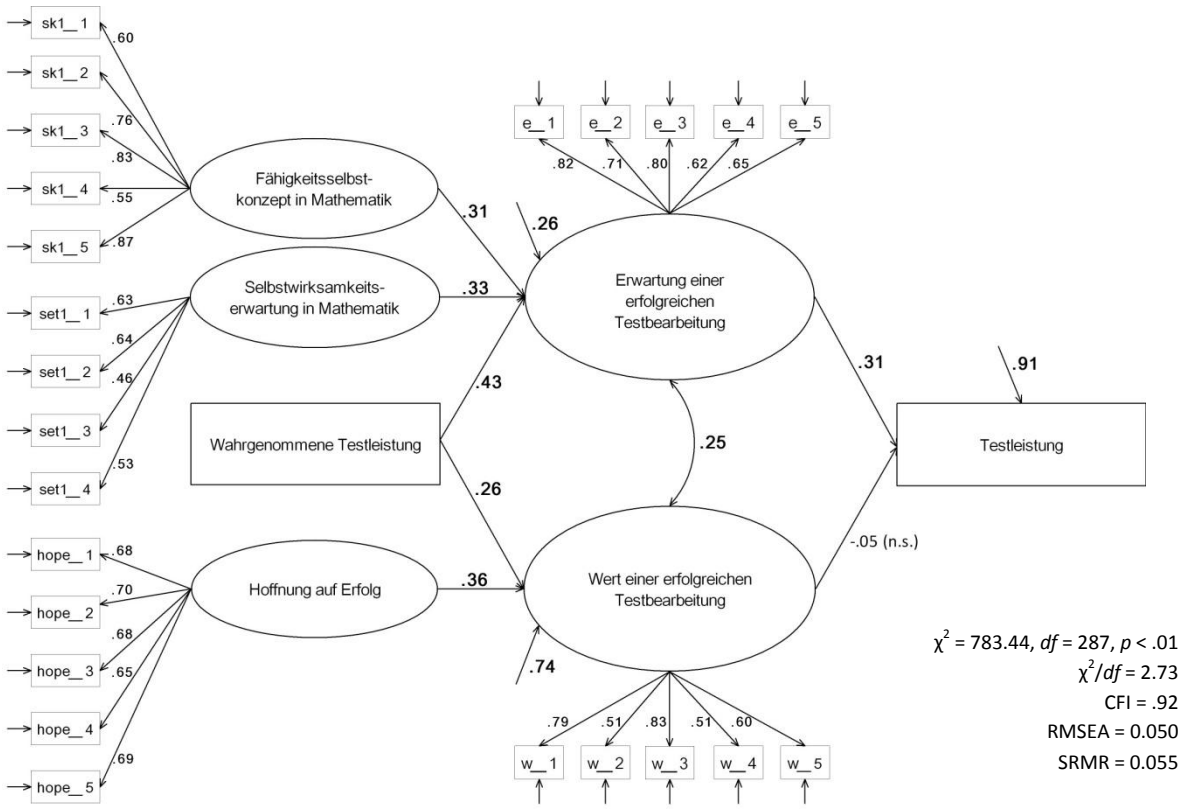



*Abbildung 11.3*: Mess- und Strukturmodell zur Motivation zur Testbearbeitung während der ersten Teststunde (vollstandardisierte Lösung; signifikante Koeffizienten sind fett gedruckt; *N* = 703).

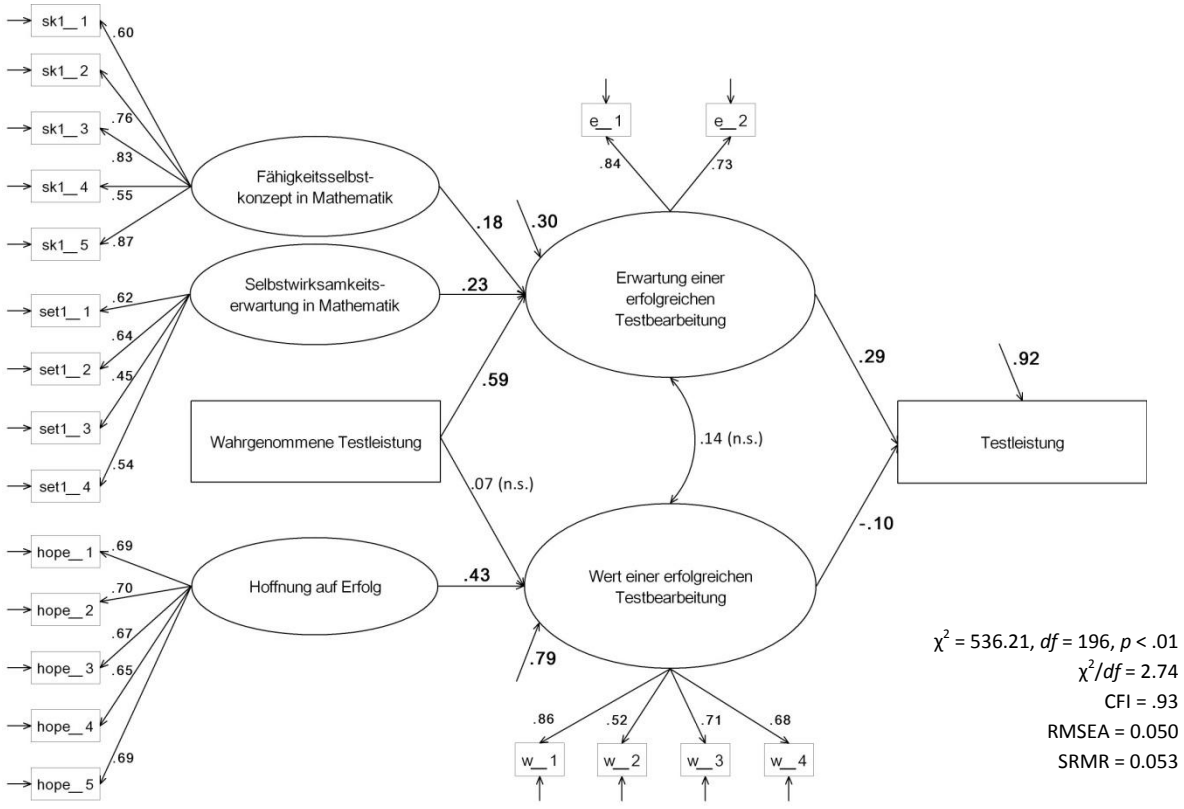



*Abbildung 11.4*: Mess- und Strukturmodell zur Motivation zur Testbearbeitung nach der ersten Teststunde (vollstandardisierte Lösung; signifikante Koeffizienten sind fett gedruckt; *N* = 703).

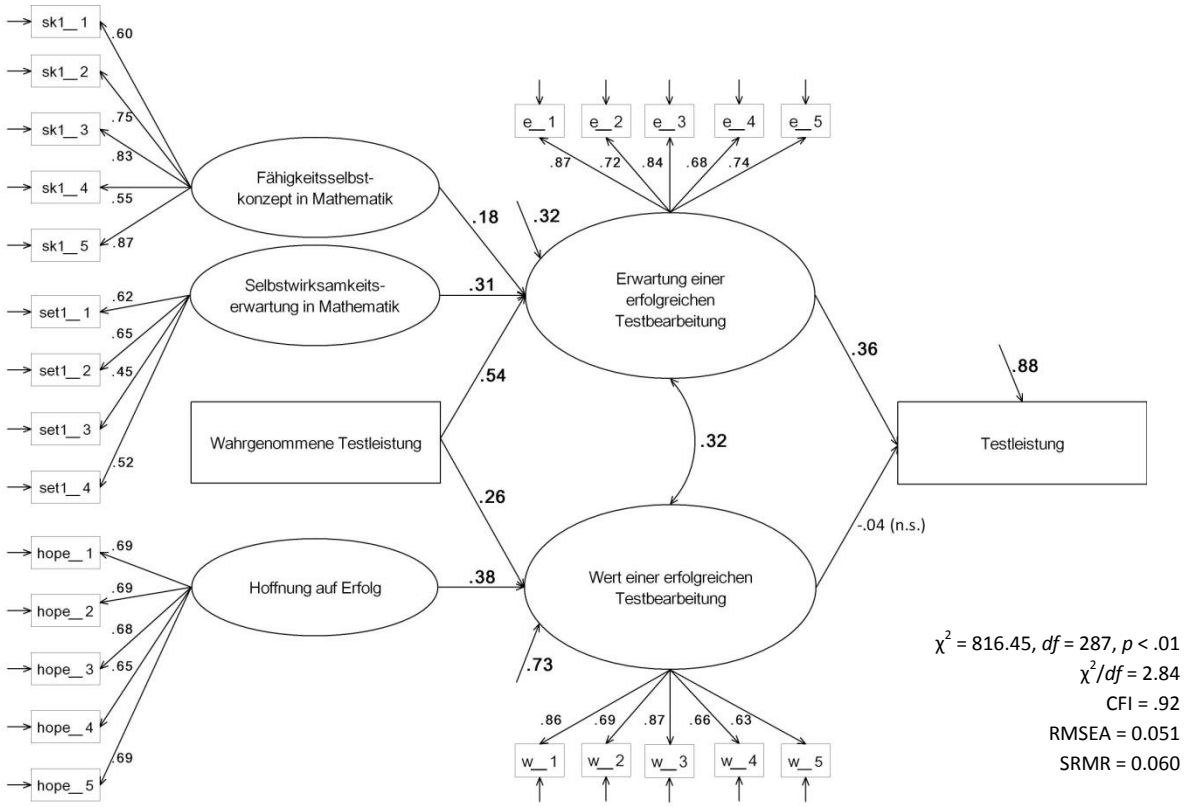



*Abbildung 11.5*: Mess- und Strukturmodell zur Motivation zur Testbearbeitung während der zweiten Teststunde (vollstandardisierte Lösung; signifikante Koeffizienten sind fett gedruckt).

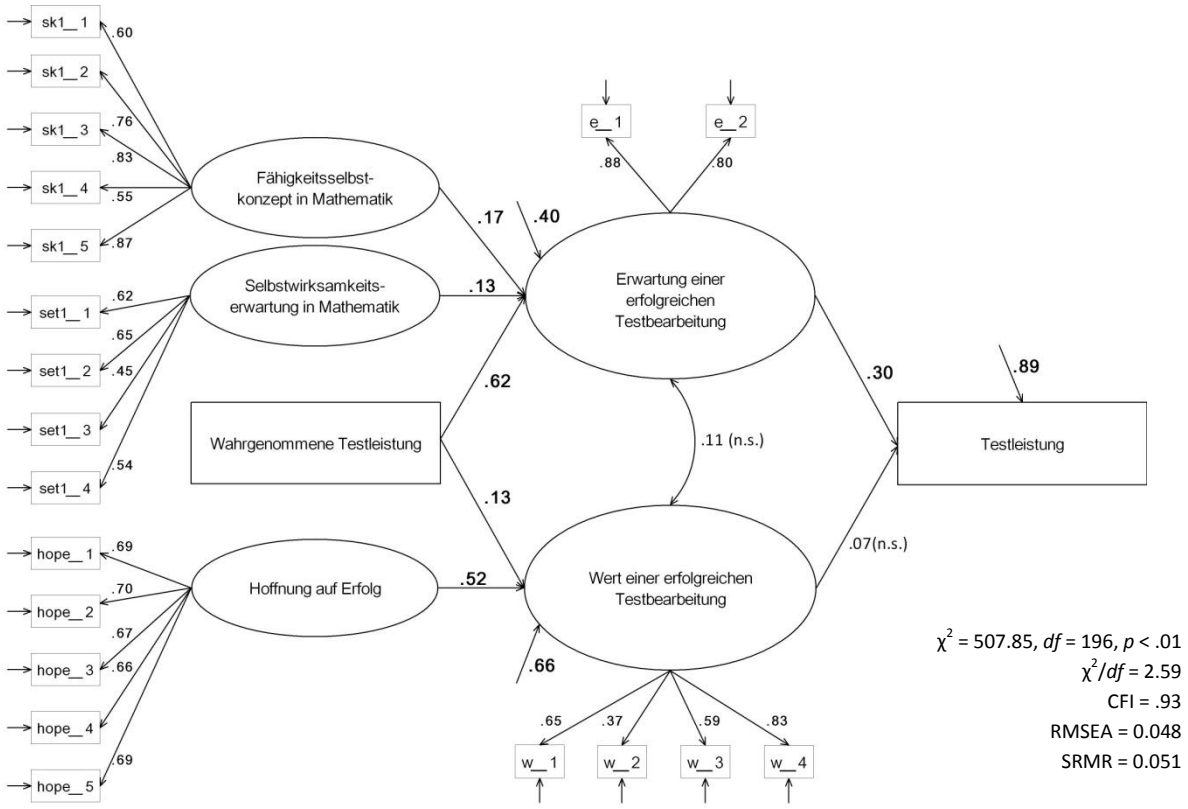



*Abbildung 11.6*: Mess- und Strukturmodell zur Motivation zur Testbearbeitung nach der zweiten Teststunde (vollstandardisierte Lösung; signifikante Koeffizienten sind fett gedruckt; *N* = 703).

*Tabelle 11.1:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.49 | 0.02 | 154.78 | < .01 | 2.34 | 0.02 | 111.56 | < .01 |
| SK ($\gamma_{01}$) | 0.41 | 0.03 | 11.73 | < .01 | 0.32 | 0.04 | 7.49 | < .01 |
| H ($\gamma_{02}$) | 0.25 | 0.03 | 8.73 | < .01 | 0.28 | 0.04 | 6.88 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.02 | -3.40 | < .01 | -0.06 | 0.03 | -2.40 | <. 05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.00 | 0.01 | 0.14 | .89 | 0.02 | 0.02 | 0.80 | .43 |
| SK ($\gamma_{21}$) | -0.02 | 0.02 | -0.79 | .43 | 0.01 | 0.03 | 0.16 | .87 |
| H ($\gamma_{22}$) | -0.04 | 0.03 | -1.13 | .26 | -0.05 | 0.04 | -1.09 | .28 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .38; nach dem Test: *ICC* = .59; $R^2$ = .24;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.2:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für Personen mit hoher Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.67 | 0.03 | 87.91 | < .01 | 2.53 | 0.04 | 61.88 | < .01 |
| SK ($\gamma_{01}$) | 0.42 | 0.07 | 6.50 | < .01 | 0.44 | 0.06 | 7.05 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.01 | 0.03 | -0.21 | .83 | -0.03 | 0.04 | -0.77 | .44 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.03 | 0.02 | -1.62 | .11 | -0.01 | 0.03 | -0.40 | .69 |
| SK ($\gamma_{21}$) | 0.01 | 0.04 | 0.24 | .81 | -0.08 | 0.04 | -2.11 | <. 05 |

*Anmerkungen.* Während des Tests: *ICC* = .77; $R^2$ = .29; nach dem Test: *ICC* = .63; $R^2$ = .20;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.3:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für Personen mit geringer Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.16 | 0.04 | 48.71 | < .01 | 2.02 | 0.08 | 27.02 | < .01 |
| SK ($\gamma_{01}$) | 0.39 | 0.07 | 5.57 | < .01 | 0.26 | 0.09 | 2.94 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.25 | 0.08 | -3.27 | < .01 | -0.24 | 0.07 | -3.20 | < .01 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.08 | 0.06 | 1.34 | .18 | 0.11 | 0.07 | 1.44 | .15 |
| SK ($\gamma_{21}$) | 0.06 | 0.05 | 1.22 | .22 | 0.11 | 0.12 | 0.94 | .35 |

*Anmerkungen.* Während des Tests: *ICC* = .60; $R^2$ = .32; nach dem Test: *ICC* = .41; $R^2$ = .17;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.4:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für die Hauptschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.47 | 0.05 | 52.52 | < .01 | 2.39 | 0.06 | 38.19 | < .01 |
| SK ($\gamma_{01}$) | 0.28 | 0.07 | 4.25 | < .01 | 0.17 | 0.07 | 2.33 | .02 |
| H ($\gamma_{02}$) | 0.36 | 0.06 | 5.59 | < .01 | 0.37 | 0.08 | 4.76 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.12 | 0.04 | -2.95 | < .01 | -0.15 | 0.06 | -2.64 | < .01 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.01 | 0.04 | 0.32 | .75 | 0.00 | 0.06 | 0.00 | > .99 |
| SK ($\gamma_{21}$) | 0.14 | 0.06 | 2.21 | <. 05 | 0.07 | 0.08 | 0.90 | .37 |
| H ($\gamma_{22}$) | -0.08 | 0.07 | -1.13 | .26 | -0.11 | 0.11 | -0.99 | .32 |

*Anmerkungen.* Während des Tests: *ICC* = .72; $R^2$ = .39; nach dem Test: *ICC* = .51; $R^2$ = .21;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.5:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für die Realschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.47 | 0.03 | 83.22 | < .01 | 2.23 | 0.04 | 60.01 | < .01 |
| SK ($\gamma_{01}$) | 0.39 | 0.05 | 7.49 | < .01 | 0.37 | 0.05 | 8.19 | < .01 |
| H ($\gamma_{02}$) | 0.20 | 0.05 | 3.75 | < .01 | 0.22 | 0.06 | 3.63 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.03 | -2.47 | < .05 | -0.04 | 0.04 | -1.05 | .29 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.02 | 0.03 | -0.59 | .55 | 0.06 | 0.04 | 1.74 | .08 |
| SK ($\gamma_{21}$) | -0.04 | 0.04 | -1.00 | .32 | -0.02 | 0.05 | -0.48 | .63 |
| H ($\gamma_{22}$) | 0.00 | 0.06 | 0.02 | .98 | -0.07 | 0.07 | -0.97 | .33 |

*Anmerkungen.* Während des Tests: *ICC* = .72; $R^2$ = .36; nach dem Test: *ICC* = .60; $R^2$ = .24;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.6:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung für das Gymnasium.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.56 | 0.03 | 91.03 | < .01 | 2.46 | 0.04 | 66.80 | < .01 |
| SK ($\gamma_{01}$) | 0.52 | 0.04 | 13.67 | < .01 | 0.37 | 0.05 | 8.01 | < .01 |
| H ($\gamma_{02}$) | 0.13 | 0.06 | 2.02 | <. 05 | 0.21 | 0.07 | 3.01 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.02 | 0.03 | -0.81 | .42 | -0.03 | 0.04 | -0.87 | .39 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.01 | 0.03 | 0.47 | .64 | -0.02 | 0.04 | -0.50 | .62 |
| SK ($\gamma_{21}$) | -0.10 | 0.04 | -2.62 | < .01 | 0.02 | 0.05 | 0.43 | .67 |
| H ($\gamma_{22}$) | -0.01 | 0.05 | -0.26 | .80 | -0.05 | 0.06 | -0.71 | .48 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .44; nach dem Test: *ICC* = .61; $R^2$ = .27;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.7*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.02 | 106.84 | < .01 | 2.31 | 0.03 | 81.85 | < .01 |
| SK ($\gamma_{01}$) | 0.36 | 0.05 | 7.78 | < .01 | 0.30 | 0.05 | 5.77 | < .01 |
| H ($\gamma_{02}$) | 0.24 | 0.05 | 4.84 | < .01 | 0.27 | 0.07 | 4.15 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.02 | -3.07 | < .01 | -0.03 | 0.03 | -0.91 | .36 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.03 | 0.03 | 0.95 | .34 | 0.07 | 0.04 | 1.84 | .07 |
| SK ($\gamma_{21}$) | 0.00 | 0.03 | 0.01 | .99 | 0.01 | 0.03 | 0.17 | .86 |
| H ($\gamma_{22}$) | -0.06 | 0.05 | -1.26 | .21 | -0.03 | 0.07 | -0.47 | .64 |

*Anmerkungen.* Während des Tests: *ICC* = .71; $R^2$ = .32; nach dem Test: *ICC* = .63; $R^2$ = .22;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.8:* Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für Personen mit hoher Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.65 | 0.04 | 60.51 | < .01 | 2.45 | 0.06 | 42.54 | < .01 |
| SK ($\gamma_{01}$) | 0.30 | 0.09 | 3.27 | < .01 | 0.39 | 0.09 | 4.49 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | 0.00 | 0.03 | 0.00 | > .99 | 0.01 | 0.05 | 0.19 | .85 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.03 | 0.03 | -1.00 | .32 | 0.06 | 0.05 | 1.16 | .25 |
| SK ($\gamma_{21}$) | 0.07 | 0.04 | 1.68 | .09 | -0.07 | 0.05 | -1.39 | .17 |

*Anmerkungen.* Während des Tests: *ICC* = .79; $R^2$ = .32; nach dem Test: *ICC* = .68; $R^2$ = .17;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.9*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für Personen mit geringer Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.20 | 0.07 | 32.28 | < .01 | 2.03 | 0.10 | 20.74 | < .01 |
| SK ($\gamma_{01}$) | 0.42 | 0.08 | 5.51 | < .01 | 0.33 | 0.10 | 3.45 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | 0.14 | 0.07 | 1.89 | .06 | -0.23 | 0.10 | -2.28 | <. 05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.14 | 0.07 | 1.89 | .06 | 0.17 | 0.10 | 1.74 | .08 |
| SK ($\gamma_{21}$) | 0.09 | 0.08 | 1.07 | .29 | 0.04 | 0.08 | 0.46 | .65 |

*Anmerkungen.* Während des Tests: *ICC* = .62; $R^2$ = .34; nach dem Test: *ICC* = .47; $R^2$ = .20;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.10*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für die Hauptschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.45 | 0.06 | 40.17 | < .01 | 2.35 | 0.09 | 25.38 | < .01 |
| SK ($\gamma_{01}$) | 0.17 | 0.09 | 1.97 | < .05 | 0.06 | 0.11 | 0.57 | .57 |
| H ($\gamma_{02}$) | 0.31 | 0.13 | 2.33 | <. 05 | 0.37 | 0.15 | 2.53 | < .05 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.12 | 0.07 | -1.62 | .11 | -0.07 | 0.09 | -0.79 | .43 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.11 | 0.03 | 4.14 | < .01 | 0.07 | 0.10 | 0.68 | .49 |
| SK ($\gamma_{21}$) | 0.24 | 0.06 | 3.84 | < .01 | 0.05 | 0.09 | 0.61 | .55 |
| H ($\gamma_{22}$) | -0.07 | 0.10 | -0.69 | .49 | -0.08 | 0.11 | -0.73 | .47 |

*Anmerkungen.* Während des Tests: *ICC* = .65; $R^2$ = .30; nach dem Test: *ICC* = .62; $R^2$ = .16;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.11*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für die Realschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.51 | 0.03 | 92.65 | < .01 | 2.25 | 0.06 | 38.62 | < .01 |
| SK ($\gamma_{01}$) | 0.38 | 0.06 | 6.50 | < .01 | 0.38 | 0.06 | 6.10 | < .01 |
| H ($\gamma_{02}$) | 0.22 | 0.07 | 3.31 | < .01 | 0.20 | 0.12 | 1.68 | .09 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.09 | 0.03 | -2.65 | < .01 | -0.06 | 0.07 | -0.86 | .39 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.02 | 0.04 | -0.34 | .73 | 0.13 | 0.05 | 2.59 | < .01 |
| SK ($\gamma_{21}$) | -0.04 | 0.03 | -1.61 | .11 | 0.01 | 0.03 | 0.34 | .74 |
| H ($\gamma_{22}$) | -0.08 | 0.07 | -1.26 | .21 | -0.03 | 0.08 | -0.40 | .69 |

*Anmerkungen.* Während des Tests: *ICC* = .72; $R^2$ = .31; nach dem Test: *ICC* = .63; $R^2$ = .27;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.12*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei intransparenter Testinstruktion für das Gymnasium.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.56 | 0.04 | 62.77 | < .01 | 2.40 | 0.03 | 94.01 | < .01 |
| SK ($\gamma_{01}$) | 0.51 | 0.08 | 6.77 | < .01 | 0.36 | 0.07 | 5.23 | < .01 |
| H ($\gamma_{02}$) | 0.07 | 0.07 | 0.93 | .35 | 0.34 | 0.12 | 2.86 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.01 | 0.03 | -0.42 | .67 | 0.03 | 0.04 | 0.81 | .42 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.01 | 0.04 | 0.28 | .78 | 0.02 | 0.05 | 0.28 | .78 |
| SK ($\gamma_{21}$) | -0.13 | 0.04 | -2.92 | < .01 | -0.01 | 0.05 | -0.28 | .78 |
| H ($\gamma_{22}$) | 0.13 | 0.08 | 1.67 | .10 | -0.10 | 0.07 | -1.36 | .17 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .39; nach dem Test: *ICC* = .65; $R^2$ = .29;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.13*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.03 | 97.96 | < .01 | 2.37 | 0.03 | 74.63 | < .01 |
| SK ($\gamma_{01}$) | 0.45 | 0.04 | 11.81 | < .01 | 0.33 | 0.05 | 6.89 | < .01 |
| H ($\gamma_{02}$) | 0.26 | 0.04 | 6.23 | < .01 | 0.29 | 0.04 | 7.16 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.03 | -2.68 | < .01 | -0.10 | 0.04 | -2.55 | < .05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.02 | 0.02 | -1.01 | .31 | -0.04 | 0.03 | -1.18 | .24 |
| SK ($\gamma_{21}$) | -0.04 | 0.03 | -1.07 | .29 | 0.01 | 0.05 | 0.22 | .82 |
| H ($\gamma_{22}$) | -0.01 | 0.05 | -0.24 | .81 | -0.08 | 0.06 | -1.30 | .19 |

*Anmerkungen.* Während des Tests: *ICC* = .75; $R^2$ = .44; nach dem Test: *ICC* = .56; $R^2$ = .25;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.14*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für Personen mit hoher Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.71 | 0.04 | 68.21 | < .01 | 2.62 | 0.04 | 64.41 | < .01 |
| SK ($\gamma_{01}$) | 0.56 | 0.05 | 10.53 | < .01 | 0.50 | 0.06 | 8.22 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.03 | 0.04 | -0.74 | .46 | -0.09 | 0.06 | -1.54 | .12 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.05 | 0.04 | -1.33 | .18 | -0.10 | 0.04 | -2.41 | <. 05 |
| SK ($\gamma_{21}$) | -0.06 | 0.06 | -1.06 | .29 | -0.09 | 0.06 | -1.51 | .13 |

*Anmerkungen.* Während des Tests: *ICC* = .76; $R^2$ = .36; nach dem Test: *ICC* = .59; $R^2$ = .24;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.15*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für Personen mit geringer Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.12 | 0.09 | 24.22 | < .01 | 2.02 | 0.12 | 17.41 | < .01 |
| SK ($\gamma_{01}$) | 0.35 | 0.14 | 2.52 | < .05 | 0.17 | 0.16 | 1.08 | .28 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.22 | 0.10 | -2.07 | <. 05 | -0.25 | 0.13 | -1.94 | .05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.01 | 0.09 | -0.06 | .96 | 0.02 | 0.08 | 0.21 | .83 |
| SK ($\gamma_{21}$) | 0.01 | 0.10 | 0.12 | .90 | 0.19 | 0.23 | 0.83 | .41 |

*Anmerkungen.* Während des Tests: *ICC* = .56; $R^2$ = .30; nach dem Test: *ICC* = .32; $R^2$ = .13;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.16*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für die Hauptschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.08 | 32.60 | < .01 | 2.43 | 0.05 | 47.78 | < .01 |
| SK ($\gamma_{01}$) | 0.38 | 0.12 | 3.22 | < .01 | 0.28 | 0.10 | 2.69 | .01 |
| H ($\gamma_{02}$) | 0.42 | 0.12 | 3.42 | < .01 | 0.39 | 0.07 | 5.95 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.13 | 0.05 | -2.65 | < .01 | -0.24 | 0.09 | -2.52 | < .05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.08 | 0.04 | -1.91 | .06 | -0.06 | 0.07 | -0.84 | .40 |
| SK ($\gamma_{21}$) | 0.02 | 0.09 | 0.25 | .81 | 0.07 | 0.07 | 1.03 | .30 |
| H ($\gamma_{22}$) | -0.02 | 0.08 | -0.22 | .83 | -0.11 | 0.09 | -1.22 | .22 |

*Anmerkungen.* Während des Tests: *ICC* = .78; $R^2$ = .52; nach dem Test: *ICC* = .40; $R^2$ = .30;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.17*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für die Realschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.43 | 0.04 | 70.25 | < .01 | 2.22 | 0.05 | 47.13 | < .01 |
| SK ($\gamma_{01}$) | 0.41 | 0.05 | 8.42 | < .01 | 0.36 | 0.07 | 5.11 | < .01 |
| H ($\gamma_{02}$) | 0.18 | 0.04 | 4.10 | < .01 | 0.26 | 0.07 | 3.80 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.04 | 0.03 | -1.39 | .16 | -0.03 | 0.05 | -0.62 | .53 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.02 | 0.04 | -0.46 | .64 | -0.01 | 0.05 | -0.20 | .84 |
| SK ($\gamma_{21}$) | -0.05 | 0.05 | -0.95 | .34 | -0.06 | 0.08 | -0.80 | .43 |
| H ($\gamma_{22}$) | 0.10 | 0.09 | 1.11 | .27 | -0.10 | 0.10 | -1.06 | .29 |

*Anmerkungen.* Während des Tests: *ICC* = .71; $R^2$ = .41; nach dem Test: *ICC* = .56; $R^2$ = .22;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.18*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei transparenter Testinstruktion für das Gymnasium.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.57 | 0.04 | 61.97 | < .01 | 2.52 | 0.06 | 44.93 | < .01 |
| SK ($\gamma_{01}$) | 0.53 | 0.05 | 11.30 | < .01 | 0.38 | 0.08 | 4.74 | < .01 |
| H ($\gamma_{02}$) | 0.18 | 0.07 | 2.52 | < .01 | 0.13 | 0.07 | 1.97 | < .05 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.04 | 0.04 | -0.94 | .35 | -0.09 | 0.04 | -2.01 | <. 05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.01 | 0.04 | 0.30 | .76 | -0.05 | 0.06 | -0.76 | .45 |
| SK ($\gamma_{21}$) | -0.06 | 0.06 | -1.07 | .29 | 0.06 | 0.09 | 0.62 | .53 |
| H ($\gamma_{22}$) | -0.13 | 0.08 | -1.55 | .12 | -0.04 | 0.08 | -0.47 | .64 |

*Anmerkungen.* Während des Tests: *ICC* = .74; $R^2$ = .48; nach dem Test: *ICC* = .58; $R^2$ = .27;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.19*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.02 | 109.11 | < .01 | 2.36 | 0.03 | 82.61 | < .01 |
| SK ($\gamma_{01}$) | 0.41 | 0.04 | 11.38 | < .01 | 0.32 | 0.04 | 7.56 | < .01 |
| TS ($\gamma_{02}$) | -0.01 | 0.04 | -0.16 | .88 | -0.03 | 0.04 | -0.62 | .53 |
| H ($\gamma_{03}$) | 0.23 | 0.03 | 9.04 | < .01 | 0.26 | 0.04 | 7.19 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.02 | -3.30 | < .01 | -0.06 | 0.03 | -2.26 | <. 05 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .38; nach dem Test: *ICC* = .59; $R^2$ = .24;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit; H: Hoffnung auf Erfolg.

*Tabelle 11.20*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für Personen mit hoher Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.65 | 0.04 | 75.79 | < .01 | 2.51 | 0.05 | 46.44 | < .01 |
| SK ($\gamma_{01}$) | 0.42 | 0.06 | 6.77 | < .01 | 0.40 | 0.07 | 6.17 | < .01 |
| TS ($\gamma_{02}$) | 0.01 | 0.06 | 0.16 | .87 | 0.03 | 0.07 | 0.37 | .71 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.01 | 0.03 | -0.19 | .85 | -0.03 | 0.04 | -0.68 | .50 |

*Anmerkungen.* Während des Tests: *ICC* = .77; $R^2$ = .29; nach dem Test: *ICC* = .63; $R^2$ = .20;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit.

*Tabelle 11.21*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für Personen mit geringer Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.24 | 0.06 | 40.36 | < .01 | 2.14 | 0.07 | 32.58 | < .01 |
| SK ($\gamma_{01}$) | 0.42 | 0.06 | 6.79 | < .01 | 0.33 | 0.07 | 4.44 | < .01 |
| TS ($\gamma_{02}$) | -0.09 | 0.08 | -1.13 | .26 | -0.16 | 0.09 | -1.75 | .08 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.23 | 0.07 | -3.13 | < .01 | -0.22 | 0.09 | -2.55 | < .05 |

*Anmerkungen.* Während des Tests: *ICC* = .60; $R^2$ = .32; nach dem Test: *ICC* = .41; $R^2$ = .17;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit.

*Tabelle 11.22*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für die Hauptschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.42 | 0.07 | 33.93 | < .01 | 2.32 | 0.06 | 37.12 | < .01 |
| SK ($\gamma_{01}$) | 0.36 | 0.10 | 3.77 | < .01 | 0.22 | 0.10 | 2.15 | <. 05 |
| TS ($\gamma_{02}$) | 0.13 | 0.10 | 1.25 | .21 | 0.15 | 0.08 | 1.92 | .05 |
| H ($\gamma_{03}$) | 0.33 | 0.09 | 3.85 | < .01 | 0.34 | 0.09 | 3.86 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.13 | 0.05 | -2.49 | < .05 | -0.16 | 0.06 | -2.67 | < .01 |

*Anmerkungen.* Während des Tests: *ICC* = .72; $R^2$ = .39; nach dem Test: *ICC* = .51; $R^2$ = .22;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit; H: Hoffnung auf Erfolg.

*Tabelle 11.23*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für die Realschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.50 | 0.03 | 95.59 | < .01 | 2.32 | 0.04 | 52.76 | < .01 |
| SK ($\gamma_{01}$) | 0.37 | 0.04 | 9.69 | < .01 | 0.36 | 0.04 | 8.52 | < .01 |
| TS ($\gamma_{02}$) | -0.07 | 0.06 | -1.22 | .22 | -0.13 | 0.06 | -2.03 | <. 05 |
| H ($\gamma_{03}$) | 0.20 | 0.03 | 7.76 | < .01 | 0.19 | 0.05 | 3.66 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.07 | 0.03 | -2.38 | <. 05 | -0.03 | 0.04 | -0.76 | .45 |

*Anmerkungen.* Während des Tests: *ICC* = .72; $R^2$ = .37; nach dem Test: *ICC* = .60; $R^2$ = .25;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit; H: Hoffnung auf Erfolg.

*Tabelle 11.24*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei adaptivem Testalgorithmus für das Gymnasium.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.56 | 0.04 | 70.97 | < .01 | 2.46 | 0.05 | 52.35 | < .01 |
| SK ($\gamma_{01}$) | 0.48 | 0.03 | 14.57 | < .01 | 0.38 | 0.04 | 8.58 | < .01 |
| TS ($\gamma_{02}$) | 0.01 | 0.05 | 0.10 | .92 | -0.01 | 0.03 | -0.40 | .69 |
| H ($\gamma_{03}$) | 0.12 | 0.08 | 1.58 | .12 | 0.19 | 0.07 | 2.55 | < .05 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.02 | 0.02 | -0.80 | .42 | -0.03 | 0.03 | -1.00 | .32 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .44; nach dem Test: *ICC* = .61; $R^2$ = .27;
SK: Fähigkeitsselbstkonzept in Mathematik; TS: Testschwierigkeit; H: Hoffnung auf Erfolg.

*Tabelle 11.25*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei mittlerer Testschwierigkeit für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.47 | 0.02 | 108.33 | < .01 | 2.35 | 0.03 | 82.58 | < .01 |
| SK ($\gamma_{01}$) | 0.41 | 0.04 | 10.45 | < .01 | 0.33 | 0.04 | 7.36 | < .01 |
| H ($\gamma_{02}$) | 0.24 | 0.04 | 5.58 | < .01 | 0.24 | 0.06 | 4.35 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.04 | 0.02 | -1.99 | < .05 | -0.10 | 0.03 | -2.86 | < .01 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.03 | 0.02 | 1.26 | .21 | 0.04 | 0.03 | 1.53 | .13 |
| SK ($\gamma_{21}$) | -0.04 | 0.03 | -1.20 | .23 | -0.05 | 0.05 | -1.07 | .29 |
| H ($\gamma_{22}$) | -0.08 | 0.05 | -1.62 | .11 | -0.05 | 0.08 | -0.64 | .52 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .37; nach dem Test: *ICC* = .55; $R^2$ = .23;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.26*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei hoher Testschwierigkeit für die Gesamtstichprobe.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.52 | 0.03 | 83.62 | < .01 | 2.33 | 0.03 | 68.00 | < .01 |
| SK ($\gamma_{01}$) | 0.40 | 0.05 | 8.62 | < .01 | 0.31 | 0.05 | 6.42 | < .01 |
| H ($\gamma_{02}$) | 0.26 | 0.02 | 6.42 | < .01 | 0.33 | 0.06 | 5.99 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.10 | 0.03 | -3.61 | < .01 | -0.04 | 0.03 | -1.19 | .24 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.03 | 0.02 | -1.54 | .13 | -0.01 | 0.04 | -0.24 | .81 |
| SK ($\gamma_{21}$) | 0.02 | 0.04 | 0.58 | .56 | 0.06 | 0.04 | 1.53 | .13 |
| H ($\gamma_{22}$) | 0.01 | 0.05 | 0.24 | .81 | -0.05 | 0.06 | -0.77 | .44 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .41; nach dem Test: *ICC* = .64; $R^2$ = .26;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.27*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei mittlerer Testschwierigkeit für Personen mit hoher Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.66 | 0.04 | 75.93 | < .01 | 2.55 | 0.06 | 44.65 | < .01 |
| SK ($\gamma_{01}$) | 0.40 | 0.08 | 4.72 | < .01 | 0.44 | 0.07 | 6.37 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | 0.00 | 0.04 | 0.05 | .96 | -0.09 | 0.05 | -1.89 | .06 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.03 | 0.03 | -0.85 | .40 | -0.03 | 0.05 | -0.60 | .55 |
| SK ($\gamma_{21}$) | -0.01 | 0.06 | -0.16 | .88 | -0.15 | 0.06 | -2.62 | < .01 |

*Anmerkungen.* Während des Tests: *ICC* = .77; $R^2$ = .25; nach dem Test: *ICC* = .60; $R^2$ = .16;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.28*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei mittlerer Testschwierigkeit für Personen mit geringer Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.10 | 0.06 | 37.13 | < .01 | 2.17 | 0.11 | 20.19 | < .01 |
| SK ($\gamma_{01}$) | 0.44 | 0.13 | 3.38 | < .01 | 0.32 | 0.17 | 1.92 | .06 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.24 | 0.08 | -2.93 | < .01 | -0.44 | 0.11 | -3.91 | < .01 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.21 | 0.06 | 3.49 | < .01 | 0.07 | 0.10 | 0.65 | .52 |
| SK ($\gamma_{21}$) | 0.12 | 0.11 | 1.05 | .30 | 0.22 | 0.20 | 1.12 | .26 |

*Anmerkungen.* Während des Tests: *ICC* = .63; $R^2$ = .42; nach dem Test: *ICC* = .24; $R^2$ = .32;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.29*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei hoher Testschwierigkeit für Personen mit hoher Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.67 | 0.05 | 54.81 | < .01 | 2.51 | 0.05 | 48.52 | < .01 |
| SK ($\gamma_{01}$) | 0.44 | 0.08 | 5.72 | < .01 | 0.45 | 0.08 | 5.93 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.02 | 0.04 | -0.47 | .64 | 0.01 | 0.04 | 0.31 | .76 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.04 | 0.03 | -1.25 | .21 | -0.01 | 0.04 | -0.11 | .91 |
| SK ($\gamma_{21}$) | 0.03 | 0.04 | 0.75 | .46 | -0.03 | 0.06 | -0.58 | .56 |

*Anmerkungen.* Während des Tests: *ICC* = .78; $R^2$ = .34; nach dem Test: *ICC* = .67; $R^2$ = .25;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.30*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei hoher Testschwierigkeit für Personen mit geringer Hoffnung auf Erfolg.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.24 | 0.08 | 29.39 | < .01 | 1.87 | 0.08 | 22.68 | < .01 |
| SK ($\gamma_{01}$) | 0.36 | 0.11 | 3.14 | < .01 | 0.23 | 0.08 | 2.72 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.28 | 0.11 | -2.57 | < .01 | -0.06 | 0.08 | -0.75 | <. 05 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.06 | 0.09 | -0.71 | .48 | 0.15 | 0.11 | 1.36 | .18 |
| SK ($\gamma_{21}$) | 0.06 | 0.09 | 0.68 | .50 | 0.04 | 0.10 | 0.39 | .70 |

*Anmerkungen.* Während des Tests: *ICC* = .57; $R^2$ = .27; nach dem Test: *ICC* = .55; $R^2$ = .10;
SK: Fähigkeitsselbstkonzept in Mathematik.

*Tabelle 11.31*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei mittlerer Testschwierigkeit für die Hauptschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.42 | 0.11 | 22.71 | < .01 | 2.36 | 0.07 | 32.37 | < .01 |
| SK ($\gamma_{01}$) | 0.31 | 0.10 | 3.01 | < .01 | 0.29 | 0.15 | 1.96 | < .05 |
| H ($\gamma_{02}$) | 0.07 | 0.09 | 0.74 | .46 | 0.27 | 0.09 | 2.95 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.04 | 0.04 | -1.05 | .29 | -0.20 | 0.06 | -3.44 | < .01 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.14 | 0.07 | 1.90 | .06 | 0.10 | 0.04 | 2.23 | <. 05 |
| SK ($\gamma_{21}$) | 0.21 | 0.08 | 2.53 | < .05 | -0.05 | 0.11 | -0.49 | .63 |
| H ($\gamma_{22}$) | 0.04 | 0.13 | 0.32 | .75 | -0.13 | 0.18 | -0.74 | .46 |

*Anmerkungen.* Während des Tests: *ICC* = .78; $R^2$ = .36; nach dem Test: *ICC* = .50; $R^2$ = .18;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.32*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei mittlerer Testschwierigkeit für die Realschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.53 | 0.03 | 73.74 | < .01 | 2.29 | 0.04 | 56.20 | < .01 |
| SK ($\gamma_{01}$) | 0.32 | 0.09 | 3.40 | < .01 | 0.37 | 0.05 | 7.05 | < .01 |
| H ($\gamma_{02}$) | 0.30 | 0.10 | 3.18 | < .01 | 0.22 | 0.08 | 2.81 | .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.08 | 0.07 | -1.14 | .25 | -0.05 | 0.06 | -0.79 | .43 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.03 | 0.04 | 0.56 | .58 | 0.05 | 0.04 | 1.29 | .20 |
| SK ($\gamma_{21}$) | -0.06 | 0.06 | -1.03 | .30 | -0.07 | 0.06 | -1.10 | .27 |
| H ($\gamma_{22}$) | -0.13 | 0.08 | -1.62 | .11 | -0.10 | 0.12 | -0.86 | .39 |

*Anmerkungen.* Während des Tests: *ICC* = .70; $R^2$ = .35; nach dem Test: *ICC* = .58; $R^2$ = .25;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.33*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei mittlerer Testschwierigkeit für das Gymnasium.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.47 | 0.04 | 57.55 | < .01 | 2.44 | 0.06 | 40.59 | < .01 |
| SK ($\gamma_{01}$) | 0.54 | 0.06 | 8.94 | < .01 | 0.29 | 0.06 | 4.87 | < .01 |
| H ($\gamma_{02}$) | 0.04 | 0.10 | 0.39 | .70 | 0.19 | 0.08 | 2.35 | <. 05 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | 0.03 | 0.04 | 0.83 | .40 | -0.07 | 0.07 | -1.11 | .27 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.11 | 0.06 | 1.69 | .09 | 0.01 | 0.05 | 0.30 | .77 |
| SK ($\gamma_{21}$) | -0.16 | 0.08 | -1.99 | < .05 | -0.00 | 0.06 | -0.05 | .96 |
| H ($\gamma_{22}$) | 0.10 | 0.11 | 0.86 | .39 | -0.03 | 0.09 | -0.39 | .70 |

*Anmerkungen.* Während des Tests: *ICC* = .69; $R^2$ = .41; nach dem Test: *ICC* = .51; $R^2$ = .20;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.34*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei hoher Testschwierigkeit für die Hauptschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b/SE* | *p* | *b* | *SE* | *b/SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.51 | 0.07 | 37.39 | < .01 | 2.42 | 0.09 | 27.12 | < .01 |
| SK ($\gamma_{01}$) | 0.25 | 0.09 | 2.83 | < .01 | 0.10 | 0.07 | 1.40 | .16 |
| H ($\gamma_{02}$) | 0.48 | 0.07 | 6.73 | < .01 | 0.51 | 0.08 | 6.58 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.18 | 0.07 | -2.51 | < .05 | -0.10 | 0.10 | -1.04 | .30 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.06 | 0.06 | -1.09 | .28 | -0.09 | 0.06 | -1.63 | .10 |
| SK ($\gamma_{21}$) | 0.14 | 0.09 | 1.63 | .10 | 0.15 | 0.13 | 1.16 | .25 |
| H ($\gamma_{22}$) | -0.12 | 0.06 | -2.03 | <. 05 | -0.13 | 0.14 | -0.88 | .38 |

*Anmerkungen.* Während des Tests: *ICC* = .64; $R^2$ = .44; nach dem Test: *ICC* = .52; $R^2$ = .28;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.35*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei hoher Testschwierigkeit für die Realschule.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.45 | 0.04 | 57.62 | < .01 | 2.18 | 0.07 | 33.36 | < .01 |
| SK ($\gamma_{01}$) | 0.43 | 0.06 | 7.86 | < .01 | 0.40 | 0.07 | 5.97 | < .01 |
| H ($\gamma_{02}$) | 0.18 | 0.07 | 2.57 | < .01 | 0.23 | 0.09 | 2.55 | < .05 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.10 | 0.03 | -3.22 | < .01 | -0.05 | 0.05 | -1.03 | .30 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | 0.00 | 0.03 | 0.11 | .91 | 0.07 | 0.05 | 1.43 | .15 |
| SK ($\gamma_{21}$) | -0.02 | 0.04 | -0.43 | .67 | 0.01 | 0.08 | 0.14 | .89 |
| H ($\gamma_{22}$) | 0.08 | 0.09 | 0.86 | .39 | -0.04 | 0.09 | -0.43 | .67 |

*Anmerkungen.* Während des Tests: *ICC* = .73; $R^2$ = .39; nach dem Test: *ICC* = .62; $R^2$ = .25;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

*Tabelle 11.36*: Mehrebenen-Wachstumsmodell zur Vorhersage der Erfolgserwartung bei hoher Testschwierigkeit für das Gymnasium.

| | Während des Tests | | | | Nach dem Test | | | |
|---|---|---|---|---|---|---|---|---|
| Feste Effekte | *b* | *SE* | *b*/*SE* | *p* | *b* | *SE* | *b*/*SE* | *p* |
| Für das Intercept $\beta_{0j}$ | | | | | | | | |
| Intercept ($\gamma_{00}$) | 2.63 | 0.05 | 56.25 | < .01 | 2.48 | 0.04 | 59.47 | < .01 |
| SK ($\gamma_{01}$) | 0.51 | 0.05 | 9.78 | < .01 | 0.44 | 0.07 | 5.99 | < .01 |
| H ($\gamma_{02}$) | 0.15 | 0.09 | 1.58 | .12 | 0.31 | 0.10 | 3.02 | < .01 |
| Für den Messzeitpunkt ($\beta_{1j}$) | | | | | | | | |
| Intercept ($\gamma_{10}$) | -0.05 | 0.03 | -2.17 | <. 05 | 0.01 | 0.04 | 0.30 | .76 |
| Für den Testalgorithmus ($\beta_{2j}$) | | | | | | | | |
| Intercept ($\gamma_{20}$) | -0.04 | 0.02 | -2.15 | <. 05 | -0.05 | 0.06 | -0.83 | .41 |
| SK ($\gamma_{21}$) | -0.04 | 0.04 | -1.20 | .23 | 0.06 | 0.04 | 1.58 | .11 |
| H ($\gamma_{22}$) | 0.08 | 0.08 | 1.01 | .31 | -0.07 | 0.10 | -0.71 | .48 |

*Anmerkungen.* Während des Tests: *ICC* = .78; $R^2$ = .48; nach dem Test: *ICC* = .71; $R^2$ = .38;
SK: Fähigkeitsselbstkonzept in Mathematik; H: Hoffnung auf Erfolg.

# Tabellarischer Lebenslauf

## Persönliche Daten

| | |
|---|---|
| Name: | Regine Asseburg |
| Geburtsdatum, -ort: | 19.04.1978, Hamburg |
| Nationalität: | Deutsch |

## Akademischer Werdegang und berufliche Tätigkeiten

| | |
|---|---|
| 07/2011 | Promotion an der Philosophischen Fakultät der Christian-Albrechts-Universität zu Kiel (Fach: Psychologie) |
| Seit 08/2006 | Wissenschaftliche Mitarbeiterin am Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel |
| 03/2006-07/2006 | Mitarbeiterin im Test-Lektorat der Hogrefe Verlag GmbH & Co. KG, Göttingen |
| 09/2005 | Abschluss des Studiengangs Diplom-Psychologie an der Ruprecht-Karls-Universität Heidelberg; Titel der Diplomarbeit: „Untersuchungen zur Fairness des IST 2000 R und des CFT 3“; Erstgutachter: Prof. Dr. Manfred Amelang, Zweitgutachter: Prof. Dr. Hans-Joachim Ahrens |
| 10/2001-02/2002 | Auslandssemester an der Université de Fribourg (Schweiz) |
| 09/2001 | Abschluss der Vordiplom-Prüfungen in Psychologie, Ruprecht-Karls-Universität Heidelberg |
| 10/1999 – 09/2005 | Psychologie-Studium (Diplom) an der Ruprecht-Karls-Universität Heidelberg |
| 10/1998 – 09/1999 | Germanistik-, Mathematik- und Theologie-Studium (Gymnasiallehramt) an der Ruprecht-Karls-Universität Heidelberg und der Carl-von-Ossietzky-Universität Oldenburg |

## Schulbildung

| | |
|---|---|
| 06/1997 | Erhalt der Allgemeinen Hochschulreife am Gymnasium Tostedt |

Kiel, im Sommer 2011 Regine Asseburg