Wie aus den Daten ersichtlich wird, liegt das DIC für jeden einzelnen Untertest beim 2PL-Modell mit Zeitdiskrimination (also beim vollen Modell) am niedrigsten. Der Unterschied zwischen dem vollen Modell und dem nächst schlechteren Modell (1PL-Modell mit Zeitdiskrimination) beträgt in jedem Fall über 100. Das Vollmodell kann deshalb als das am besten passende IRT-Modell betrachtet werden.

Die Modelltests belegen eindrucksvoll, dass die Bearbeitungszeit (mithin die Leseflüssigkeit) bei der Erfassung der Lesekompetenz keinesfalls unberücksichtigt bleiben darf. Vielmehr stellt sie einen zentralen Parameter der Lesefähigkeit dar.

Bei der Itemselektion wurde zum einen berücksichtigt, dass genügend schwierige Items im Test vertreten sein sollten, zum anderen sollten alle berechneten Trennschärfen für ein Item (also Trennschärfen nach KTT und nach IRT) eine Mindesthöhe erreichen. Diese beiden Kriterien führten zum Aussortieren von 8 Items beim Wortverständnistest, von 2 Items beim Satzverständnistest und von 5 Items beim Textverständnistest. Die verbliebenen Items weisen alle mindestens Trennschärfen von 0.2 nach KTT und von 0.35 nach IRT auf.

Anschließend wurden die Aufgaben neu gereiht. Diese vorläufige Neureihung fand für den Satz- und Textverständnistest auf der Basis von Schwierigkeit und Bearbeitungsdauer statt. Beim Wortverständnis wird allerdings die Bearbeitungsdauer nicht nur durch die Lesegeschwindigkeit des Probanden, sondern auch durch die Position des Targets unter den Antwortalternativen beeinflusst. Deshalb berücksichtigten wir beim Wortverständnistest zur Reihung anstatt des Zeitverbrauchs zusätzlich die Silbenzahl des Targets.

5.3.3 DIF-Analyse

Ein Test wird im Allgemeinen dann als fair betrachtet, wenn er Testpersonen nicht aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit systematisch diskriminiert (Testkuratorium der Föderation deutscher Psychologenverbände, 1986). Dies bedeutet nicht notwendigerweise, dass z. B. Jungen und Mädchen in einem Test im Mittel gleich gut abschneiden müssen. Sie müssen nur dann gleich gut abschneiden, wenn sie auch gleiche latente Fähigkeiten im gemessenen Konstrukt aufweisen. Um die Testfairness jedes einzelnen Items in Bezug auf das Geschlecht und den Migrationshintergrund zu überprüfen, wurden deshalb „differential item functioning"-Analysen (DIF-Analysen) durchgeführt. Da diese Analysen zum Ausschluss verschiedener Items führten, beschreiben wir sie im vorliegenden Kapitel, obwohl es sich bei den DIF-Analysen eigentlich um Fragen der Testfairness handelt, welche in Kapitel 6.4 thematisiert wird.

Um die DIF-Analysen durchführen zu können, wurde für jedes Item ein Leistungsmaß berechnet (Itemscore[2] geteilt durch Bearbeitungszeit). Anschließend wurde überprüft, ob bei einer linearen Regression das Geschlecht sowie die Interaktion aus Geschlecht und Testrohwert über den reinen Testrohwert hinaus noch Varianz an diesem Leistungsmaß aufklären. Als Testrohwert wurde hierbei diejenige Anzahl an Items aus dem Wort-, Satz- bzw. Textverständnisteil herangezogen, die die Probanden jeweils in der limitierten Bearbeitungszeit und bei Vorgabe der neuen Itemreihenfolge richtig gelöst hätten. Äquivalente DIF-Analysen wurden zur Untersuchung der Unterschiede von Kindern mit und ohne Migrationshintergrund durchgeführt. Kinder ohne Migrationshintergrund erhielten dabei das Rating „0", Kinder mit einem eingewanderten Elternteil erhielten das Rating „1" und Kinder mit zwei eingewanderten Elternteilen das Rating „2". Bei beiden Analysen ist allerdings zu beachten, dass sie aufgrund der metrischen Daten und der hohen Fallzahl eine sehr hohe Power besitzen und außerdem aufgrund der hohen Itemzahl eine α-Inflation vorliegt. Um die Effekte von Geschlecht und Migrationshintergrund (uniforme DIFs) sowie der jeweiligen Interaktion mit dem Testrohwert (nonuniforme DIFs) nicht zu überschätzen, wurden nur solche Items aus dem Test ausgeschlossen, bei denen sich über 1 % der Itemvarianz durch die zusätzliche Aufnahme der Variable Geschlecht oder Migrationshintergrund bzw. der entsprechenden Interaktion in die Regressionsgleichung aufklären ließ. Allerdings gab es im Wort-, Satz- und Textverständnistest insgesamt nur 4 Items, bei denen dies der Fall war. Diese Items wurden aus dem Test entfernt. Die verbliebenen Effekte von Migrationshintergrund und Geschlecht sind klein. Zudem heben sich die DIFs unterschiedlicher Items in den Testrohwerten größtenteils gegenseitig auf. Der ELFE II-Test kann deshalb unseres Erachtens als fair hinsichtlich Geschlecht und Migrationshintergrund betrachtet werden. Im Kapitel 6.4 („Testfairness") werden zusätzlich grafische Modelltests der Testfairness berichtet, die nach der endgültigen Itemselektion durchgeführt wurden.

Nach der Selektion aller Items wurden die Modellparameter erneut mit cirt (Klein Entink, 2013) geschätzt. Eine Übersicht über die resultierenden Parameter findet sich in Anhang D.

5.3.4 Lokale stochastische Unabhängigkeit

Damit in der IRT-Skalierung die Itemparameter korrekt geschätzt werden und damit Skalen sich als eindimensional erweisen können, darf die Wahrscheinlichkeit, mit der ein Item gelöst wird, nicht davon abhängen, ob vorher irgendein anderes Items richtig gelöst wurde (lokale stochastische

2 „0" für „nicht gelöst" und „1" für „gelöst"

Unabhängigkeit). Eine Verletzung dieser Annahme führt in der Regel zu einer Überschätzung der Testgüte. Da beim Textverständnistest zu den meisten Texten mehr als eine Frage gestellt wird und aus diesem Grund eine Verletzung der Unabhängigkeit gegeben sein könnte, ist bei ELFE II diese Prüfung besonders interessant. Die Unabhängigkeit der Beantwortung der Aufgaben innerhalb der Subtests wurde nach der Itemselektion und der Neuberechnung der Parameter anhand der Vorgehensweise von Sinharay, Johnson und Stern (2006) mittels CIRT überprüft. Weder im Satz- noch Textverständnistest ergaben sich Hinweise auf die Verletzung der Unabhängigkeitsannahme. Im Wortverständnistest ergaben sich bei drei Itempaaren („15. Stuhl" und „75. Stadion"; „42. Detektiv" und „52. Bilderrahmen"; „56. Papagei" und „63. Schublade") Hinweise auf Abhängigkeiten. Da bei diesem Untertest bei 75 Aufgaben die Anzahl an Einzelvergleichen mit 2775 sehr groß ist, war allerdings zu erwarten, dass einige der Vergleiche „zufällig" signifikant ausfallen. Es gibt außerdem keine inhaltlich sinnvolle Erklärung dafür, wieso das richtige Erkennen des Wortes „Detektiv" Einfluss auf die Lösung beim Wort „Bilderrahmen" haben sollte (ebenso bei „Papagei" und „Schublade"). Wir gehen deshalb davon aus, dass hier keine relevante Einschränkung des Testverfahrens vorliegt.

5.4 Itemselektion bei der Schwellenmessung

Bei der Schwellenmessung wurde festgelegt, dass ein Item nur dann in den endgültigen Test aufgenommen werden sollte, wenn es von mindestens 90 % aller Kinder bei einer Darbietungsdauer von 5 Sekunden richtig kategorisiert wird.

Zusätzlich wurden auch für die Items dieses Testteils DIF-Analysen durchgeführt. Da wir wegen der Darbietung ohne relevante Geschwindigkeitsbegrenzung im Gegensatz zu Wort-, Satz- und Textverständnis keine Leistungsmaße berechnen konnten, wurde hierfür nur die binäre Antwortkodierung („richtig" oder „falsch") herangezogen. Die Analysen wurden deshalb mit binär-logistischer Regression durchgeführt. Für 9 der Items ergaben sich dabei signifikante Einflüsse von Geschlecht oder Migrationshintergrund. Diese Items wurden aussortiert.

Zusammen mit dem oben definierten 90 %-Kriterium wären nun nach dem Aussortieren nur 11 einsilbige Items im Test verblieben. Da von jeder Sorte aber mindestens 12 Items benötigt wurden, wurde ein einsilbiges Item („Mais") in den Test aufgenommen, obwohl dieses Item nur von 89 % aller Kinder richtig kategorisiert worden war. Ein differenzieller Effekt von Geschlecht oder Migrationshintergrund lag bei diesem Item nicht vor. Die im Test verbliebenen Items wurden im Mittel von 95 % aller Probanden richtig beantwortet (die mit KTT berechneten Schwierigkeiten aller Items der Schwellenmessung finden sich in Anhang D).

Die Items wurden schließlich für die Testrevision in 12 Blöcken à drei Items gruppiert. Jeder Block enthält ein einsilbiges, ein zweisilbiges sowie ein dreisilbiges Wort.

Auf eine Modellierung von Testparametern nach IRT wurde beim Lesegeschwindigkeitstest verzichtet. Es scheint zwar zunächst so, als ob gerade bei diesem adaptiven Test die IRT die geeignete Testtheorie zur Beschreibung der Itemkennwerte wäre. Allerdings muss dabei berücksichtigt werden, dass der eigentliche Leistungsparameter, der dem Test zugrunde liegt, nämlich die minimale Darbietungszeit, die zum Lesen eines Wortes benötigt wird, in der Pilotierung noch gar nicht variiert werden konnte.

5.5 Beschreibung der Testrevision

5.5.1 Testinhalt und -durchführung

Die revidierte Testversion enthält insgesamt 75 Wortverständnisitems, 36 Satzverständnisitems und 26 Textverständnisitems. Der nur in der Computerform enthaltene Test zur Schwellenmessung der Worterkennung enthält 36 Items.

Im Rahmen der Testrevision wurde auch untersucht, für welche Darbietungszeiten des Wort-, Satz- und Textverständnistests die geringsten Boden- und Deckeneffekte auftreten würden. Als optimal erwiesen sich hierbei trotz der nun höheren Itemzahlen die bereits in der früheren Testfassung verwendeten Darbietungszeiten (3 Minuten für den Wortverständnistest, 3 Minuten für den Satzverständnistest und 7 Minuten für den Textverständnistest).

Da sich der Test allerdings über einen großen Entwicklungszeitraum der Lesefähigkeit erstreckt, stellt dieser reguläre Darbietungsmodus trotzdem für einige Kinder der unteren Klassenstufen eine sehr hohe und für Kinder der hohen Klassenstufen eine relative niedrige Anforderung dar. Es gibt deshalb in ELFE II die Möglichkeit, verschiedene Kurzversionen durchzuführen, um die Testanforderungen noch besser an diese einzelnen Probandengruppen anzupassen. Den Decken- und Bodeneffekten wurde also nicht nur auf Ebene der Aufgabenkonstruktion, sondern auch auf Ebene der Testdurchführung (und Normierung) Rechnung getragen.

Die folgenden zwei Kurzversionen wurden in ELFE II neu eingeführt:

Tabelle 6.1
Odd-Even-Split-Half-Reliabilität von ELFE II nach Klassenstufe, Testform und Testversion

Klassenstufe	*n*	Wort[1]	Satz[1]	Text[1, 3]	Gesamtergebnis	Kurzversion 1–3[4]
Papier: Standardversion und Kurzversion 1–3						
1	234	.97	.89	.71 (.83)	.95	.95
2	282	.98	.95	.87 (.93)	.97	.98
3	295	.98	.97	.87 (.93)	.98	.99
4	309	.99	.96	.82 (.90)	.97	
5	200	.98	.94	.78 (.88)	.95	
6	146	.97	.94	.76 (.86)	.96	
7	54	.96	.92	.73 (.84)	.91	
Gesamt[5]	**1 520**	**.98**	**.94**	**.80 (.89)**	**.96**	**.98**
Computer: Standardversion und Kurzversion 1–3						
1	140	.95 (.94)	.86 (.92)	.63 (.77)	.92	.95
2	217	.95 (.97)	.90 (.93)	.77 (.87)	.95	.96
3	264	.96 (.97)	.95 (.96)	.80 (.89)	.96	.97
4	282	.93 (.96)	.92 (.94)	.79 (.88)	.95	
5	141	.91 (.94)	.93 (.94)	.79 (.88)	.93	
6	100	.92 (.95)	.95 (.96)	.80 (.89)	.95	
7	34	.82 (.90)	.85 (.92)	.75 (.86)	.91	
Gesamt[5]	**1 178**	**.93 (.95)**	**.91 (.94)**	**.76 (.87)**	**.94**	**.96**
Computer: Kurzversion 4–7						
4	282	.93	.91	.81 (.89)	.96	
5	137	.90	.89	.80 (.89)	.93	
6	99	.95	.94	.79 (.89)	.94	
7	34	.96	.93	.83 (.91)	.96	
Gesamt[5]	**552**	**.94**	**.92**	**.81 (.89)**	**.95**	

Anmerkungen: Alle Korrelationen signifikant mit $p < .001$ (zweiseitig). [1] Zur Berechnung der Korrelationen wurde der jeweilige Rohwert herangezogen (d. h. die Anzahl richtig bearbeiteter Items). [2] Korrelation der Leistungsmaße (Anzahl der richtig gelösten Items geteilt durch die logarithmierte Zeit) in Klammern. [3] Nach Spearman-Brown aufgewertete Korrelation in Klammern. [4] Die Kurzversion 1–3 besteht nur aus Wort- und Satzverständnisteil. [5] Die Berechnung der mittleren Korrelationen über alle Klassenstufen hinweg erfolgte über Fishers z-Transformation.

Tabelle 6.2
Retestkorrelationen, Intraklassenkorrelationen und absolute Retesteffekte

	Rohwerte			Normwerte (*T*-Werte)		
	r_{tt}[1]	*ICC*	d_{Cohen}	r_{tt}[2]	*ICC*	d_{Cohen}[3]
Wortverständnis	.85	.96	0.23	.83	.88	0.24
Satzverständnis	.91	.97	0.19	.90	.93	0.20
Textverständnis	.85	.94	0.31	.81	.86	0.34
Gesamtergebnis				**.93**	**.93**	**0.30**

Anmerkung: Alle Korrelationen signifikant mit $p < .001$ (zweiseitig). [1] Die Berechnung der Retestkorrelationen r_{tt} erfolgte für die Rohwerte unter Auspartialisierung der Beschulungsdauer. [2] Da die Normwerte bereits um den Effekt der Beschulungsdauer bereinigt sind, wird hier keine zusätzliche Korrektur benötigt. [3] Bei der Berechnung von d_{Cohen} wurde für die Normwerte eine Standardabweichung von 10 Punkten (*T*-Wertskala) zugrunde gelegt.

Die Ergebnisse der Wiederholungsmessung zeigen für alle Untertests eine gute bis sehr gute um die Beschulungsdauer bereinigte Retestkorrelation. Die Ergebnisse für die Normwerte liegen nur minimal niedriger. Der Gesamttest weist eine hervorragende Retestkorrelation auf. Auch die Intraklassenkorrelationen bewegen sich durchweg im hohen bis sehr hohen Bereich. Die absolute Veränderung nach einem Monat beträgt bei den Untertests zwischen 2.0 und 3.4 *T*-Wertpunkten, beim Gesamttest 3.0 *T*-Wertpunkte. Eine solche Abweichung liegt von der Größenordnung her im Bereich der normalen Messunsicherheit. (Anm.: Die Größe des 90 %-Konfidenzintervalls beträgt beispielsweise beim Gesamttest 6.7 *T*-Wertpunkte.)

Die Schwellenmessung der Worterkennung stellt einen Sonderfall in der Testkonstruktion dar, da sie erstens nur in der Computerform vorkommt und dieser Test zweitens kriterial ausgerichtet ist: Kinder sollen spätestens bis zum Ende der vierten Klasse weniger als 200 ms zum Lesen eines Wortes benötigen. In der Regel wird diese Schwelle allerdings schon wesentlich früher erreicht. Aufgrund der Tatsache, dass ab Klassenstufe 3 bei diesem Test starke Deckeneffekte in dem Sinne auftreten, dass sehr viele Kinder die von der Software vorgegebene minimale Schwelle von 150 ms erreichen, sind die Korrelationen zwangsläufig stark eingeschränkt. Trotz dieser methodischen Limitierungen zeigte sich auch hier eine zumindest zufriedenstellende Retestkorrelation von $r_{tt} = .78$. Die Intraklassenkorrelation zeigte eine gute Übereinstimmung beider Messungen (siehe Tab. 6.3).

Tabelle 6.3
Retest- und Intraklassenkorrelationen der Schwellenmessung der Worterkennung

	r_{tt}	*ICC*
Schwellenmessung	.78	.87

Anmerkung: Alle Korrelationen signifikant mit $p < .001$ (zweiseitig). In die Berechnung flossen die Daten jener 92 Kinder ein, die im Prä- und Posttest mit der Computerform gearbeitet hatten. Vor Berechnung aller Korrelationen wurden die erzielten Schwellen logarithmiert, um die Daten zu normalisieren (vgl. auch Kap. 5.3.1).

Zusätzlich zur Erfassung der Reliabilität mittels der Retestmethode wurde die longitudinale Messinvarianz des Tests mittels latenter Strukturgleichungsmodelle untersucht. Messinvarianzprüfungen erfolgen i. d. R., indem sukzessive Parameter der Modelle restringiert werden (z. B. Ladungsmuster, Ladungen, Intercepts und Fehlervarianzen; vgl. Meredith, 1993) und der Reihe nach durch Vergleich der Modelle geprüft wird, welche Auswirkungen die Restriktionen auf die Modellpassung haben. Führt die jeweils zusätzlich eingeführte Restriktion zu einer signifikanten Verschlechterung der Modellpassung, so muss davon ausgegangen werden, dass die restringierten Parameter nicht invariant zwischen den verschiedenen Messzeitpunkten waren. Diese Überprüfung ist notwendig, da bei Wiederholungsmessungen auf Gruppenebene die Ergebnisse nur dann miteinander verglichen werden dürfen, wenn mindestens metrische Invarianz vorliegt. Das untersuchte Messmodell umfasste für jeden der beiden Zeitpunkte die Rohwerte der Untertests als Indikatoren eines latenten Faktors „Leseverständnis" und der latente Faktor von Zeitpunkt 2 wurde auf jenen von Zeitpunkt 1 regrediert. Die Schätzung erfolgte mittels des Maximum-Likelihood-Schätzers unter Lavaan (Rosseel, 2012, 0.5.20) und der Methode longInvariance der Bibliothek semTools (Pornprasertmanit et al., 2016, 0.4–11). Zur Entscheidung über die Modellpassung wurde auf die von Cheung und Rensvold (2002) vorgeschlagene Prozedur zurückgegriffen, nach der ein Unterschied im CFI der Modelle von 0.1 oder mehr eine substanzielle Verschlechterung der Modellpassung anzeigt. Die Prüfung der longitudinalen Messinvarianz ergab für den Gesamttest strikte Messinvarianz. Das bedeutet, dass nicht nur die Faktorenstruktur, sondern auch die Intercepts und Residualvarianzen der latenten Faktoren zwischen den Messungen invariant sind. Gruppenvergleiche auf der Basis von Wiederholungsmessungen mit dem Gesamttest sind folglich zulässig und absolute Unterschiede können im Sinne von Unterschieden im Personenmerkmal interpretiert werden.

6.2.3 Paralleltestreliabilität und Äquivalenz der Testformen

Zwar sind alle Aufgaben und Darbietungszeiten des Wort-, Satz- und Textverständnistests in der Papier- und der Computerform absolut identisch. Dennoch kann nicht von vorneherein davon ausgegangen werden, dass beide Testformen komplett identische Messergebnisse liefern. Schließlich gibt es verschiedene Unterschiede in der Testdurchführung. So fehlt beispielsweise in der Computerform die Möglichkeit, zu einer vorherigen Aufgabe zurückzuspringen, das Umblättern entfällt und die Markierung der richtigen Lösung nimmt unter Umständen eine andere Zeitdauer in Anspruch. Die Darbietungsform kann auch generell die Art und Weise beeinflussen, wie Aufgaben bearbeitet werden. So zeigen ältere Forschungsarbeiten, dass am Computer Aufgaben schneller, aber auch ungenauer bearbeitet werden (z. B. van de Vijver & Harsveldt, 1994). Dieser Effekt betrifft vor allem weniger komplexes Aufgabenmaterial, während höhere Verständnisleistungen davon relativ unbeeinträchtigt bleiben (Horton & Lovitt, 1994). In jüngerer Zeit hat die Debatte um Effekte des Darbietungsmediums durch die rasante Steigerung der Verfügbarkeit digitaler Medien erneut Fahrt aufgenommen. Insbesondere unter Erwachsenen und älteren Schülern wurde untersucht, ob Leseprozesse oder Arbeitsverhalten in Testsituationen durch Präsentation am Computer bedeutsam beeinflusst werden. Die meisten dieser

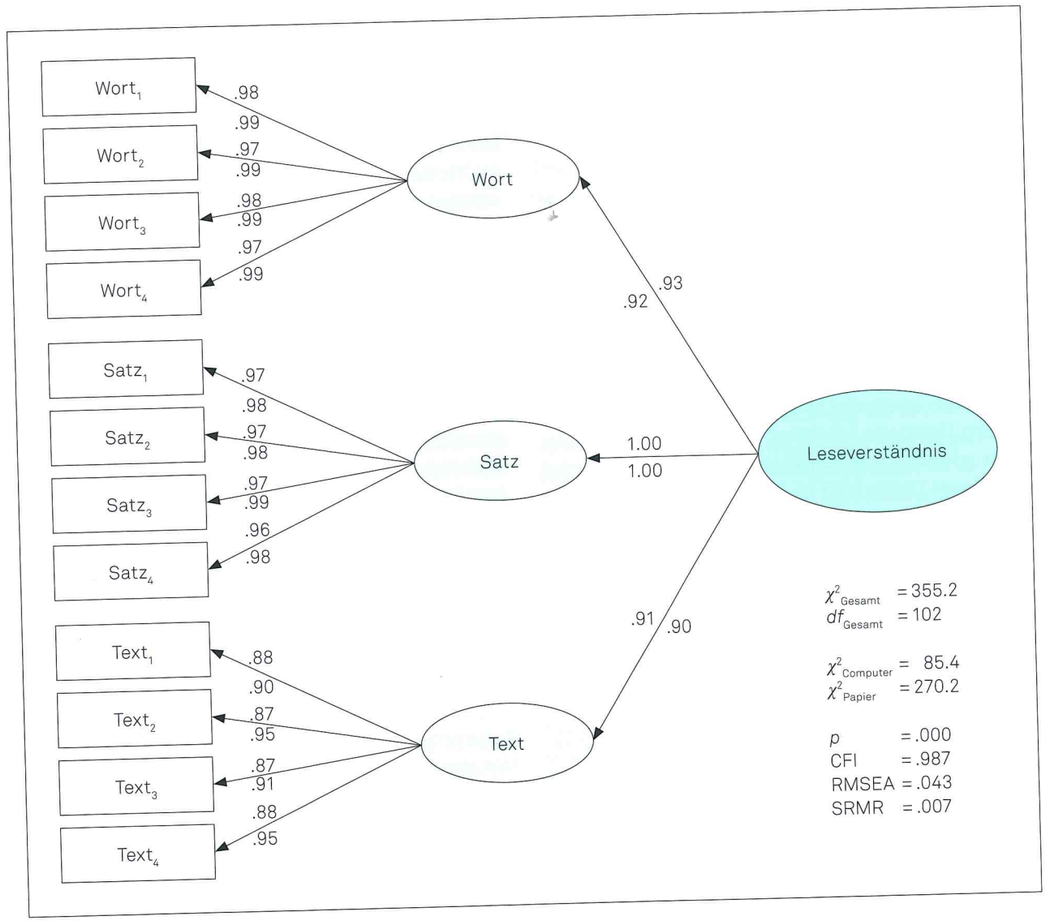

Abbildung 6.1
Darstellung des Modells der konfirmatorischen Faktorenanalyse auf der Basis der Einzelaufgabenergebnisse (Multigruppenmodell; die oberen Koeffizenten entstammen dem Modell der Papierform, die unteren der Computerform). Die Berechnung erfolgte mit Lavaan 0.5.20 unter R 3.2.5 mittels des ML-Schätzers.

Bei der Bewertung der sehr hohen Korrelationen muss außerdem auch berücksichtigt werden, dass das Modell über alle Altersbereiche gerechnet und zudem um die Messfehler bereinigt wurde.

6.3.3 Kriteriumsbezogene Validität

Die *kriteriumsbezogene Validität* gibt an, wie gut Testergebnisse mit Außenkriterien übereinstimmen. Sie kann sich etwa darauf beziehen, wie gut zukünftige Leistungen vorhergesagt werden können *(prognostische Validität)* oder – wie im Folgenden berichtet – wie gut das Verfahren mit anderen Verfahren und Informationsquellen übereinstimmt.

Konvergente und diskriminante Validität

Die *konvergente Validität* bezieht sich auf die Übereinstimmung mit Testergebnissen oder Variablen, die das gleiche Konstrukt oder ähnliche Konstrukte abbilden oder die theoriebegründet stark miteinander zusammenhängen. Hohe Korrelationen signalisieren eine Übereinstimmung oder relative Entsprechung der Konzepte. Korrelationen mit Testergebnissen oder Variablen, die davon unterscheidbare Konstrukte abbilden, sollten hingegen im Sinne *diskriminanter Validität* niedriger ausfallen. Zu berücksichtigen ist hierbei, dass intellektuelle Leistungen (vor allem solche, die an repräsentativen Stichproben erhoben wurden) immer positiv miteinander korrelieren (vgl. Rost, 2009, Kap. 2 und 4), sodass sich die Korrelationen verschiedener sprachlicher und schulischer Leistungen oft nur graduell voneinander unterscheiden.

Bereits bei ELFE 1-6 (W. Lenhard & Schneider, 2006) zeigten sich hohe Korrelationen mit einer Reihe an anderen Lesetests (z. B. WLLP, Küspert & Schneider, 1998; Knuspel-L, H. Marx, 1998) sowie mit dem Urteil der Lehrkräfte. Neben den Untersuchungen, die im Rahmen der ursprünglichen Testkonstruktion und -normierung durchgeführt wurden, liegen mittlerweile für die erste Testfassung auch etliche wissenschaftliche Studien vor, welche die Validität des Verfahrens sehr gut belegen (exemplarisch Karing, 2009; Pfost, Dörfler & Artelt, 2010; Richter, Isberner, Naumann & Kutzner, 2012; Robitzsch, Dörfler, Pfost & Artelt, 2011; Stutz, Schaffner & Schiefele, 2016).

Mit der Testrevision führten wir weitere Untersuchungen durch, um den Zusammenhang mit anderen Lesetests, dem Lehrerurteil über (schrift)sprachliche und andere schulische Leistungen, der Konzentrationsleistung und der fluiden Intelligenz zu erfassen. Im Folgenden werden diese Ergebnisse vorgestellt.

Zur Überprüfung der Übereinstimmung mit anderen Lesetests wurde exemplarisch das Salzburger Lesescreening 2-9 (SLS 2-9; Mayringer & Wimmer, 2014) herangezogen. Das SLS 2-9 besteht aus einer Liste an Sätzen, deren inhaltliche Aussagen hinsichtlich ihrer Richtigkeit (ja/nein) bewertet werden müssen. Gemessen wird, wie viele Sätze innerhalb von drei Minuten korrekt bewertet werden. Die Normierung basiert auf einer rein österreichischen Stichprobe. Die beiden Testverfahren ELFE II und SLS 2-9 wurden jeweils innerhalb der gleichen Schulstunde in Schulklassen der Klassenstufen 1 bis 4 einer baden-württembergischen Grundschule durchgeführt. Ein Vergleich der Ergebnisse fand sowohl auf Ebene der Rohwerte als auch der Normwerte statt. Es ergaben sich dabei gute Übereinstimmungen beider Testverfahren (siehe Tab. 6.5). Betrachtet man die Korrelationen der Rohwerte über alle Klassenstufen hinweg, so zeigt das SLS 2-9 eine signifikant höhere Korrelation mit dem Satzverständnistest von ELFE II im Vergleich zum Wortverständnistest, $z = 1.63$, $p = .05$, und zum Textverständnistest, $z = 1.77$, $p < .05$. Da das SLS 2-9 nur mit Sätzen arbeitet, liefert dieses Ergebnis zusätzlich einen weiteren Hinweis auf die Konstruktvalidität von ELFE II.

Zur Ermittlung der konvergenten und diskriminanten Validität von ELFE II mit verschiedenen Lehrerurteilen wurden im Rahmen der Testnormierung an mehreren Schulen zusätzlich Lehrerurteile erhoben. An einer ersten Studie nahmen 154 Kinder der Klassenstufen 1 bis 4 einer inklusiven Grundschule in Nordrhein-Westfalen teil (Klassenstufe 1: $n = 41$; Klassenstufe 2: $n = 37$; Klassenstufe 3: $n = 38$; Klassenstufe 3: $n = 38$). Die Lehrer bewerteten hier jeweils auf einer 5-stufigen Skala die Fähigkeiten im Lesen sowie die Fähigkeiten in Bezug auf gesprochene Sprache. An einer zweiten Studie nahmen 113 Kinder der Klassenstufen 2 bis 5 an Grund- und Hauptschulen in Niedersachsen teil (Klassenstufe 2: $n = 24$; Klassenstufe 3: $n = 25$; Klassenstufe 4: $n = 42$; Klassenstufe 5: $n = 23$). Hier bewerteten Lehrer außer den Lesefähigkeiten auch noch die Fähigkeiten im Sprechen, Schreiben und Rechnen auf einer 10-stufigen Skala. Aufgrund der unterschiedlichen Skalierungen in den beiden Untersuchungen werden die Ergebnisse im Folgenden getrennt berichtet. Zwecks der Vergleichbarkeit über alle Klassenstufen hinweg wurden die Ergebnisse im ELFE II Leseverständnistest vor den Analysen in Normwerte transformiert.

Tabelle 6.6 zeigt die Ergebnisse beider Untersuchungen. Die Passung des *Lehrerurteils Lesen* mit dem Gesamtergebnis von ELFE II kann mit $r \geq .7$ jeweils als sehr hoch eingestuft werden. Dagegen fielen die Korrelationen mit anderen Leistungen im Sinne der diskriminanten Validität erwartungsgemäß niedriger aus. Das Gesamtergebnis von ELFE II korrelierte in Studie 1 mit dem Lehrerurteil der allgemeinen Sprachfertigkeiten signifikant niedriger als mit dem Lehrerurteil Lesen, $z = 5.44$, $p < .001$. Gleiches konnte auch in Studie 2 in Bezug auf die allgemeinen Sprachfertigkeiten, $z = 5.43$, $p < .001$, die Schreibfertigkeiten, $z = 3.37$, $p < .001$, und die Rechenfertigkeiten, $z = 3.64$, $p < .001$, nachgewiesen werden. Dementsprechend misst ELFE II spezifisch Lese-

Tabelle 6.5
Korrelationen zwischen ELFE II und SLS 2-9

	n	Rohwerte ELFE II[2]			Normwerte ELFE II[3]			Gesamtergebnis
		Wort	Satz	Text	Wort	Satz	Text	
1. Klasse[1]	45	.498	.755	.623				
2. Klasse	27	.893	.834	.814	.893	.856	.827	.919
3. Klasse	40	.714	.738	.728	.705	.673	.656	.747
4. Klasse	34	.594	.637	.491	.598	.666	.557	.671
Gesamt	**143**	**.682**	**.742**	**.680**	**.721**	**.718**	**.676**	**.769**

Anmerkung: Alle Korrelation signifikant mit $p < .001$ (zweiseitig). [1] Da beim SLS 2-9 für die erste Klasse noch keine Normwerte verfügbar sind, entfällt dort die Berechnung der Korrelationen. [2] Die Gesamtkorrelationen wurden über alle Klassenstufen hinweg unter Auspartialisierung der Beschulungsdauer berechnet. [3] Für die Berechnung der mittleren Korrelationen über alle Klassenstufen hinweg wurden die Werte über Fishers z-Transformation gemittelt.

- **Kurzversion für erste bis dritte Klassenstufe (Kurzversion 1–3):** Der Textverständnistest stellt am Ende der ersten Klassenstufe sowie für leistungsschwache Kinder der zweiten und ggf. auch dritten Klassenstufe oft eine große Herausforderung dar. Um eine Überforderung der Testprobanden zu vermeiden, kann der Textverständnistest bei solchen Kindern deshalb auch weggelassen werden. Der Gesamtwert wird in diesen Fällen nur aus dem Wort- und dem Satzverständnisteil ermittelt. Diese Vorgehensweise ist auch dann möglich, wenn der Test zwar komplett durchgeführt wurde, aber Probleme im Textverständnistest auftraten. Auf diese Weise ist dennoch die Berechnung eines Gesamtergebnisses möglich.
- **Kurzversion für die vierte bis siebte Klassenstufe (Kurzversion 4–7):** Leistungsstarke Schüler der vierten bis siebten Klassenstufen schaffen es bisweilen, in der standardmäßig vorgegebenen Zeit alle Items des Wort-, Satz- oder Textverständnistests zu bearbeiten (dies betrifft allerdings selbst am Anfang der siebten Klassenstufe weniger als die besten 10 % einer Klassenstufe). Um auch im sehr hohen Leistungsbereich differenzieren zu können, bieten wir für die betreffenden Klassenstufen Normwerte an, die mit verkürzten Darbietungszeiten ermittelt wurden (siehe Durchführungsanleitung in Kap. 8.2).

Die jeweiligen Kurzversionen des ELFE II Leseverständnistests stellen also eine sinnvolle Alternative für besonders leistungsschwache Kinder der niedrigen oder besonders leistungsstarke Kinder der hohen Klassenstufen dar. Aufgrund der Testverkürzungen besitzen die daraus ermittelten Testergebnisse jedoch teilweise eine geringfügig niedrigere Reliabilität als die Standardversion. Zudem werden Normen für die weiterführenden Diskrepanzvergleiche zwischen Untertests und innerhalb der Untertests (siehe auch Kap. 5.5.2, Kap. 9.5 und Kap. 9.6) für die Kurzversion 4–7 nur für die Computerversion bereitgestellt. Wir empfehlen deshalb vor allem für die vierte bis siebte Klassenstufe, die Kurzversionen nur in begründeten Fällen durchzuführen. Gruppentestungen sollten generell mit der Standardversion durchgeführt werden.

Die Pseudoparalleltestversion des Tests entfällt bei der neuen Testversion komplett. Einige Anwender mögen dies bedauern. Es gibt hierfür jedoch gewichtige Gründe. Tatsächlich hängt die Bearbeitungszeit der einzelnen Items zumindest beim Wortverständnistest signifikant davon ab, an welcher Stelle unter den Antwortalternativen sich die richtige Antwort befindet. Im Allgemeinen mitteln sich zwar die unterschiedlichen Bearbeitungszeiten zwischen den beiden pseudoparallelen Testformen heraus. Vor allem bei sehr schlechten Lesern kann jedoch nicht vollständig ausgeschlossen werden, dass die unterschiedlichen Reihenfolgen der Antwortalternativen einen Einfluss auf die Leistung im Test haben. Bei Wiederholungsmessungen würde eine Pseudoparalleltestversion Retesteffekte zudem kaum reduzieren (zu Retesteffekten siehe Kap. 6.2). Lediglich bezüglich der Durchführungsobjektivität hätte die Anwendung pseudoparalleler Versionen Vorteile. Wir empfehlen deshalb, bei Testung von größeren Gruppen oder ganzen Klassen gut darauf zu achten, dass nicht vom Tischnachbarn abgeschrieben wird, beispielsweise indem jedes Kind an einem eigenen Tisch sitzt oder Sichtschirme verwendet werden.

5.5.2 Auswertung

Auch die Testauswertung wurde in ELFE II gegenüber derjenigen von ELFE 1-6 revidiert. Die Skalenwerte für die einzelnen Untertests sowie für den Gesamttest werden wie bisher aus den innerhalb der Darbietungsdauer richtig bearbeiteten Items abgeleitet. Sie stellen damit automatisch Leistungswerte gemäß der in Kapitel 5.3 beschriebenen Definition dar, da bei eingeschränkter Darbietungsdauer die Bearbeitungsgeschwindigkeit eine maßgebliche Rolle spielt. Für die Untertests werden die Ergebnisse allerdings nicht mehr in z-Werten, sondern in ganzzahligen *T*-Werten angegeben, wie dies beispielsweise auch bei anderen Schulleistungstests der Fall ist. Der Gesamtwert wird wie bisher in *T*-Werten und Prozenträngen (inkl. Konfidenzintervallen) angegeben.

Durch ein neues nonparametrisches Verfahren zur kontinuierlichen Modellierung der Normwerte (A. Lenhard, Lenhard, Suggate & Segerer, 2016) konnte die Testauswertung in zweifacher Hinsicht verbessert werden. Erstens gibt es im Gegensatz zur Vorgängerversion keine Tabellierungslücken mehr. Zweitens beschränkt sich die Anwendung des Tests nun nicht mehr nur auf Mitte und Ende eines Schuljahres, sondern der Test lässt sich zu jedem beliebigen Zeitpunkt des Schuljahres durchführen und auswerten.

Eine weitere Veränderung der Auswertung besteht darin, dass für den Wort-, Satz- und Textverständnistest Normwerte und zugehörige Auftretenshäufigkeiten für Diskrepanzen zwischen den Untertests angeboten werden. Somit kann beispielsweise ermittelt werden, ob ein Kind die Anforderung an das Wortlesen bewältigt, bei höheren Leseanforderungen auf Satz- oder Textverständnisebene aber keine der Klassenstufe angemessene Leistung mehr erbringt.

Außerdem kann für die einzelnen Untertests nicht nur die Anzahl richtig gelöster Items, sondern auch die Anzahl bearbeiteter Items ausgewertet werden. Der jeweils zugehörige *T*-Wert stellt dabei einen Indikator für den latenten Faktor Lesegeschwindigkeit ζ dar (siehe Kap. 5.3.1). In den Pilotierungsdaten korrelierte ein aus allen drei Untertests errechneter Gesamtwert für die Lesegeschwindigkeit zu $r = .92$ mit dem aus den latenten Geschwindigkeitsparametern von Wort-, Satz- und Textuntertest per Faktorenanalyse geschätzten latenten Faktor Lesegeschwindigkeit. Für sich alleine sollte die Anzahl bearbeiteter Items trotzdem nicht interpretiert werden, da sie im Einzelfall eine hohe Anzahl falscher Antworten beinhalten kann, also für sich kein gutes Leistungsmaß darstellt. Allerdings lässt die Differenz der *T*-Werte für bearbeitete und richtig gelöste Items Rückschlüsse auf den Arbeitsstil eines Probanden zu (siehe Kap. 7.5). In den Normwerttabellen wird deshalb angegeben, ob die *T*-Werte für bearbeitete und richtig gelöste Items signifikant voneinander abweichen und welcher Prozentsatz an Kindern derselben Klassenstufe eine solche oder noch höhere Abweichung zwischen bearbeiteten und richtig gelösten Items aufweist (also inwiefern eine solche Abweichung diagnostische Valenz besitzt). Somit können ggf. nützliche Förderhinweise abgeleitet werden, z. B. im Hinblick darauf, ob ein Kind eher zu langsamerem, aber dafür genauerem Lesen angehalten werden sollte, oder ob eher eine Steigerung der Lesegeschwindigkeit Priorität hat.

5.5.3 Optische und technische Neugestaltung

Auch die Testhefte sowie das Computerprogramm wurden einer Revision unterzogen. So erscheinen beim Testheft die einzelnen Antwortalternativen jetzt bei allen Untertests in hellem Grau. Sie sind somit vor allem beim Satz- und Textverständnistest leichter identifizierbar als bisher. Zwischen den letzten Items eines Untertests und der Instruktion für den nächsten Test wurde außerdem jeweils eine Seite freigelassen, damit es den Kindern nicht möglich ist, während der Instruktion eines Untertests noch vorhergehende Items zu bearbeiten. Der Auswertungsbogen wurde komplett vom Testheft entkoppelt und kann somit leichter der Schüler- oder Klientenakte zugefügt werden.

Das Computerprogramm wurde komplett neu geschrieben und an aktuelle technische Entwicklungen angepasst. Neben aktuellen Windows-Betriebssystemen (Vista, Win7, Win8 und 8.1, Win10) ist das Programm außerdem auch für Mac OS X-Nutzer verfügbar. Mithilfe des Hogrefe Datenservers können die erhobenen Daten zudem zentral auf einem Server im Netzwerk gespeichert werden.

Die Schwellenmessung der Worterkennung wurde in der bisherigen Variante direkt nach dem Wortverständnistest appliziert. Die beiden Untertests erfassen zwar sehr ähnliche Komponenten des Lesens, stehen sich also inhaltlich sehr nahe. Da die Schwellenmessung wegen der genauen zeitlichen Darbietung in der Papierform aber nicht dargeboten werden kann, stellte sie schon immer lediglich einen optionalen Test dar, der als Zusatzinformation herangezogen werden kann. Um die Darbietung der übrigen Tests zwischen Papier- und Computerform so ähnlich wie möglich zu gestalten, haben wir uns deshalb dazu entschlossen, den Test erst nach dem Textverständnistest darzubieten.

fertigkeiten und nicht allgemein schulische oder sprachliche Leistungen.

In einer weiteren Studie wurden zusätzlich zu ELFE II die Aufmerksamkeits- und Konzentrationsleistung mit dem d2-R Aufmerksamkeits- und Konzentrationstest (Brickenkamp, Schmidt-Atzert & Liepmann, 2010) und die fluide Intelligenz mit dem Grundintelligenztest Skala 1–Revision (CFT 1-R; Weiß & Osterland, 2012) erhoben. An der Untersuchung nahmen insgesamt 137 Kinder der Klassenstufen 1 bis 4 einer Grundschule in Nordrhein-Westfalen teil. 73 Kinder absolvierten die Computerform von ELFE II, 64 Kinder die Papierform. Beim d2-R müssen die Probanden innerhalb einer begrenzten Zeit so schnell wie möglich aus einer Menge an einfachen visuellen Symbolen bestimmte Zielsymbole ausfindig machen und markieren. Als Kennwert wurde die „Konzentrationsleistung" herangezogen. Bei diesem Wert spielt der Arbeitsstil des Kindes (d.h. schnell und ungenau vs. langsam und genau) keine Rolle. Der CFT 1-R stellt ein sprachfreies Maß für die fluide Intelligenz dar. Aus Ökonomiegründen wurden nur die Untertests Reihenfortsetzen, Klassifikation und Matrizen durchgeführt. Um Deckeneffekte zu vermeiden, wurden die Darbietungszeiten gegenüber der standardisierten Vorgabe verkürzt. Da deshalb keine Normwerte zugeordnet werden konnten, erfolgten alle weiteren Berechnungen mit der Rohwertsumme aus den Untertests. Um die Konzentrationsspanne der Kinder nicht zu überlasten, wurde die Erhebung an drei aufeinanderfolgenden Tagen durchgeführt, an denen aber jeweils nicht alle Kinder anwesend waren.

Tabelle 6.6
Korrelationen von ELFE II mit Lehrerurteilen im Lesen, Sprechen, Schreiben und Rechnen

	n	Normwerte ELFE II			
		Wort	Satz	Text	Gesamttest
Studie 1					
Lehrerurteil Lesen	154	.608***	.695***	.591***	**.700***
Lehrerurteil Sprechen	154	.376***	.485***	.349***	**.444***
Studie 2					
Lehrerurteil Lesen	113	.606***	.694***	.639***	**.712***
Lehrerurteil Sprechen	113	.207*	.338***	.303**	**.308***
Lehrerurteil Schreiben	113	.463***	.603***	.377***	**.528***
Lehrerurteil Rechnen	113	.396***	.530***	.440***	**.496***

Anmerkung: Alle Korrelationen signifikant (zweiseitig) mit *$p < .05$, **$p < .01$, *** $p < .001$. Die Ergebnisse wurden teilweise mit der Computer-, teilweise mit der Papierform ermittelt.

Tabelle 6.7
Korrelationen zwischen d2, CFT 1-R und ELFE II

		n	Rohwerte ELFE II[1]			Normwerte ELFE II[2]			Gesamttest
			Wort	Satz	Text	Wort	Satz	Text	
1. Klasse	d2	33	.32	.36*	–.24	.32	.34	–.20	.16
	CFT 1-R	32	.27	.32	–.20	.30	.29	–.16	.15
2. Klasse	d2	37	.52**	.59**	.33*	.52***	.59***	.31	.55***
	CFT 1-R	35	.45**	.44**	.41*	.46**	.44**	.42*	.50**
3. Klasse	d2	17	.02	.22	.30	.31	.43	.47	.40
	CFT 1-R	21	.24	.25	.29	.24	.25	.29	.11
4. Klasse	d2	35	.37*	.35*	.51**	.36*	.35*	.51**	.44**
	CFT 1-R	34	.48**	.57**	.58**	.50**	.55**	.59***	.60***
Gesamt	**d2**	**122**	**.26****	**.32*****	**.22***	**.39*****	**.47*****	**.27****	**.43*****
	CFT 1-R	**122**	**.28*****	**.32*****	**.20***	**.39*****	**.41*****	**.32*****	**.39*****

Anmerkung: Korrelationen signifikant (zweiseitig) mit * $p < .05$, ** $p < .01$, *** $p < .001$. [1] Die Gesamtkorrelationen wurden über alle Klassenstufen hinweg unter Auspartialisierung der Beschulungsdauer berechnet. [2] Für die Berechnung der mittleren Korrelationen über alle Klassenstufen hinweg wurden die Werte über Fishers z-Transformation gemittelt.

Die Ergebnisse sind in Tabelle 6.7 dargestellt. Wie erwartet, korrelieren Konzentrations- und Intelligenzleistungen niedriger mit den Ergebnissen von ELFE II als Maße der Leseleistung (z.B. SLS 2-9 oder Lehrerurteil). Die Korrelationen mit dem Gesamtergebnis von ELFE II lagen sowohl für den d2-R als auch für den CFT 1-R über alle Klassenstufen hinweg bei etwa $r = .4$. Vergleichbare Werte für die Korrelation zwischen Intelligenz und Leseleistung wurden auch in anderen Untersuchungen berichtet (z.B. Carver, 1990). Das Ergebnis stellt damit einen weiteren Beleg für die diskriminante Validität von ELFE II dar.

Kriteriumsvalidität und Subgruppenanalysen

Die Validität eines Verfahrens wird nicht zuletzt auch dadurch gestützt, dass sich Leistungsunterschiede verschiedener Personengruppen, die bereits aus der Forschung bekannt sind, in den Testergebnissen des vorliegenden Verfahrens widerspiegeln. Aufgrund des Umfangs der Analysen widmen wir diesen differenziellen Effekten ein eigenes Kapitel im Manual (siehe Kap. 7). Es finden sich dort unter anderem Analysen zu Geschlechtseffekten (Kap. 7.1), Effekten des Migrationshintergrunds (Kap. 7.2), Auswirkungen von Lese-Rechtschreibstörungen (Kap. 7.3) und der Schulform (Kap. 7.4).

6.4 Testfairness

In Kapitel 5.3.3 wurden schon itemspezifische Analysen zur Testfairness berichtet, die zum Ausschluss unfairer Items führten. An dieser Stelle möchten wir zusätzlich grafische Modelltests ergänzen, die erst nach der endgültigen Itemselektion durchgeführt wurden.

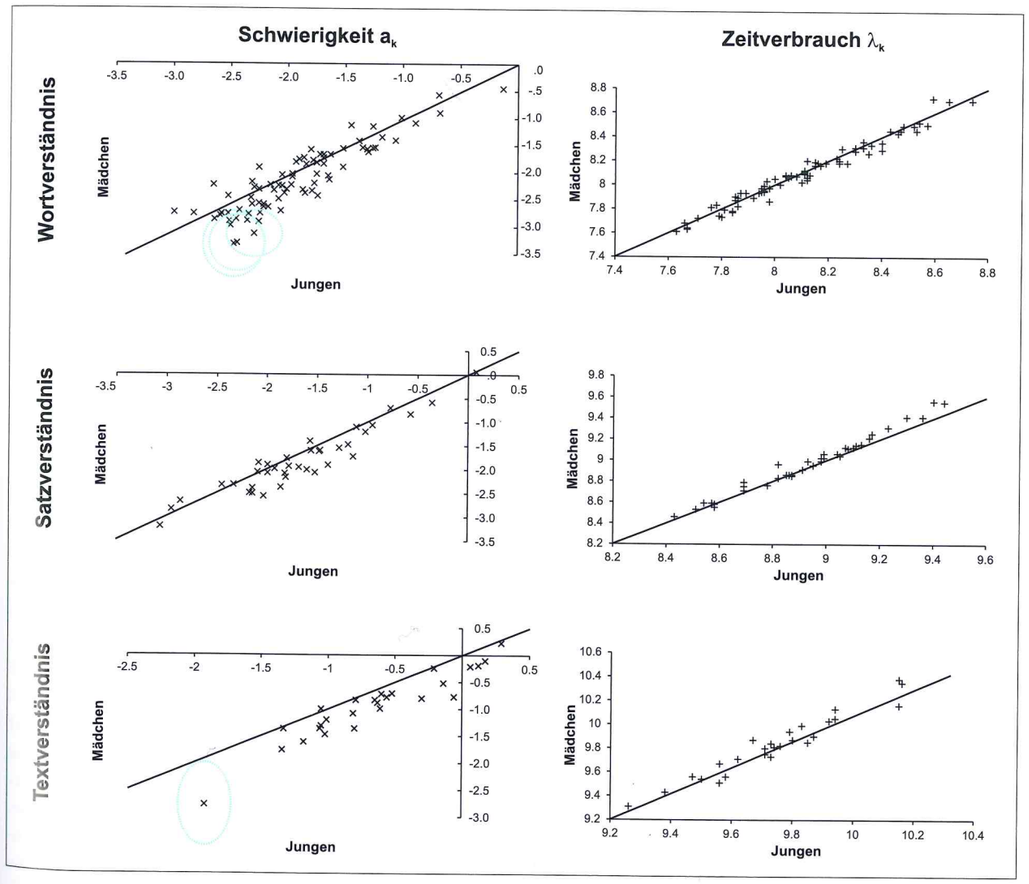

Abbildung 6.2
Grafische Modelltests des 2 × 2PL-Modells für die Parameter Itemschwierigkeit a_k und Zeitverbrauch λ_k getrennt nach Geschlecht berechnet

nterfragt werden, denn gegebenenfalls eraten. Bei Kindern der ersten drei Klasn zur Ermittlung eines reliablen Gesamtergebr die Kurzversion 1–3 zurückgegriffen werden, bei ur die Ergebnisse aus Wort- und Satzverständnistest eingehen.

Im Gegensatz zum ungenauen Arbeitsstil erhöht ein besonders genauer Arbeitsstil in der Regel die Reliabilität. Bei der Interpretation von extrem niedrigen Fehlerraten im Vergleich zur Anzahl bearbeiteter Items sollte allerdings beachtet werden, dass diese nicht unbedingt im Sinne eines Speed-Accuracy Trade-off als suboptimal langsamer Arbeitsstil interpretiert werden können. Vielmehr sind es – vor allem in den höheren Klassenstufen – die besonders schnellen und guten Leser, die extrem geringe Fehlerraten liefern, wie Abbildung 7.6 dokumentiert.

Falls ein Kind also insgesamt sehr gute Ergebnisse mit außergewöhnlich niedrigen Fehlerraten erzielt, sollte auf keinen Fall schnelleres Arbeiten empfohlen werden. Nur wenn ein Kind konsistent besonders niedrige Fehlerraten bei gleichzeitig schlechtem Ergebnis liefert, könnte dies eventuell ein Hinweis darauf sein, dass es insgesamt zu langsam arbeitet bzw. bei schnellerem Arbeiten ein besseres Ergebnis erzielen könnte. Ob in einem solchen Fall allerdings tatsächlich ein schnelleres Arbeiten empfohlen wird, sollte trotzdem sorgfältig abgewogen werden. Eventuell könnten weitere schulische Leistungsdaten zur Validierung des Ergebnisses herangezogen werden.

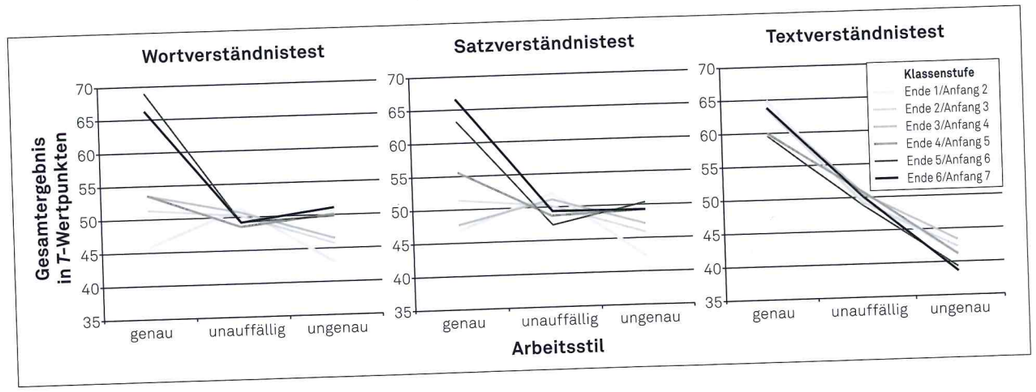

Abbildung 7.6
Gesamtergebnis in Abhängigkeit vom Arbeitsstil beim Wortverständnistest, Satzverständnistest und Testverständnistest; je dunkler die eingezeichnete Linie, desto höher die Klassenstufe

8 Testdurchführung

In diesem Kapitel finden Sie Anweisungen zur Durchführung der Papier- und Computerform des ELFE II Leseverständnistests. In Kapitel 9 wird anschließend erklärt, wie man die Rohwerte der Untertests ermittelt, in verschiedene abgeleitete Skalen umrechnet und diese interpretiert.

Bitte lesen Sie Kapitel 8 und 9 unabhängig von Ihrer bisherigen Erfahrung mit dem ELFE 1-6 Leseverständnistest sorgfältig durch, da sich verschiedene Aspekte der Testdurchführung, -auswertung und -interpretation gegenüber der früheren Ausgabe verändert haben. Um ein sicheres und interpretierbares Ergebnis zu erzielen, müssen alle Durchführungs- und Auswertungsregeln genau befolgt werden.

8.1 Anwendungszeitraum und Zielgruppe

ELFE II ist ab dem neunten Schulmonat der ersten Klassenstufe bis zum dritten Monat der siebten Klassenstufe normiert. Im Gegensatz zur Vorgängerversion liegen nicht nur Normen für Mitte und Ende des Schuljahres vor, sondern der Test kann im Normierungszeitraum zu einem beliebigen Zeitpunkt während des Schuljahres angewandt werden.

8.2 Durchführungsdauer

Der ELFE II Leseverständnistest beinhaltet in der Standardversion der Papierform den Wort- und den Satzverständnistest mit jeweils 3 Minuten Bearbeitungsdauer sowie den Textverständnistest mit 7 Minuten Bearbeitungsdauer. Bei der Computerform kommt noch der Test zur Schwellenmessung der Worterkennung dazu, dessen Durchführungsdauer von der Lese- und Reaktionsgeschwindigkeit des Kindes abhängt, in der Regel aber nicht mehr als 2 Minuten in Anspruch nimmt. Somit ergibt sich eine reine Bearbeitungsdauer von etwa 13 Minuten für die Papierform und 15 Minuten für die Computerform. Eine Gruppentestung mit der Papierform ist erfahrungsgemäß inklusive Vorbereitung, Austeilen der Testhefte, Ausfüllen der personenbezogenen Daten, Instruktion und Einsammeln der Testhefte in 20 bis 30 Minuten realisierbar.

Zusätzlich zur Standardversion können Normen, wie bereits in Kapitel 5.5.1 beschrieben, auch für zwei verschiedene Kurzversionen des ELFE II Leseverständnistests ermittelt werden:

- **Kurzversion 1–3** (für die Klassenstufen 1 bis 3): Der Textverständnistest entfällt hierbei komplett. Die Gesamtleistung wird also nur aus dem Wort- und dem Satzverständnistest ermittelt. Die reine Bearbeitungszeit verkürzt sich deshalb zu 6 Minuten (Papierform) bzw. 8 Minuten (Computerform mit Schwellenmessung der Worterkennung).
- **Kurzversion 4–7** (für die Klassenstufen 4 bis 7): Die Bearbeitungszeiten betragen dabei 2 Minuten für den Wort- und Satzverständnistest sowie 6 Minuten für den Textverständnistest. Für den Gesamttest ergeben sich also Bearbeitungszeiten von 10 Minuten (Papierform) bzw. 12 Minuten (Computerform).

Tabelle 8.1 in Kapitel 8.5.1 gibt einen Überblick über die Bearbeitungszeiten der einzelnen Untertests.

8.3 Testmaterial und -raum

Das Kind benötigt für die Testdurchführung einen aufgeräumten Arbeitsplatz. Vor Beginn der Testung ist für eine ruhige Umgebung und für angemessene Lichtverhältnisse zu sorgen. Es muss sichergestellt werden, dass das Kind während der Testung nicht durch Geräusche oder Mitschüler gestört wird.

Für die Durchführung der Testung am Computer sind außer dem Rechner keine weiteren Hilfsmittel erforderlich. Zur Durchführung der Papierform werden folgende Materialien benötigt:

- 1 Testheft pro Kind
- 2 Stifte pro Kind (davon 1 Ersatzstift). Am besten eignen sich Farbstifte, da dies die Auswertung erleichtert.
- 1 Testheft für Demonstrationszwecke für den Testleiter/die Testleiterin
- Instruktionskarte (oder Testmanual)
- Stoppuhr

Alle weiteren Hilfsmittel, insbesondere Radiergummis, Lineale, Tintenkiller, Federmäppchen sowie Hefte und sonstige Schulmaterialien müssen vor Beginn der Testung vom Arbeitsplatz entfernt werden.

8.4 Allgemeine Durchführungshinweise

Die Ergebnisse des ELFE II Leseverständnistests haben nur dann Aussagekraft, wenn der Test unter standardisierten Bedingungen durchgeführt wird. Trotz der potenziell hohen Objektivität des Verfahrens möchten wir auch noch einmal darauf hinweisen, dass es die Aufgabe des Testleiters bzw. der Testleiterin ist, für die Gleichheit von Durchführungsbedingungen für alle zu testenden Kinder zu sorgen. Dies gilt sowohl für die computerbasierte Testung als auch für die Untersuchung mit der Papierform.

Bitte beachten Sie deshalb bei der Durchführung die folgenden Punkte so genau wie möglich:

- Die Instruktionen (kursiv gedruckte Textstellen in den Kapiteln 8.5 und 8.6 bzw. auf der Instruktionskarte) sollten möglichst wortgetreu wiedergegeben werden. Zwar dürfen Füllwörter so angepasst oder ergänzt werden, dass eine möglichst natürliche mündliche Sprache entsteht. Achten Sie aber bitte genau darauf, dass hiervon keine Schlüsselwörter betroffen sind, die den Sinn der Instruktion verändern könnten.
- Instruktionen dürfen bei Nachfragen wiederholt oder in eigenen Worten formuliert werden. Es dürfen allerdings keine über die Instruktion hinausgehenden Hilfestellungen gegeben werden. Falls inhaltliche Nachfragen während der Testung kommen, können Sie zum Beispiel antworten: *„Wähle die Antwort, die am besten passt."*
- Das Kind muss mit allen Buchstaben vertraut sein. Bei Kindern, die noch nicht über sichere Kenntnis aller Buchstaben verfügen, kann die Leseverständnisleistung noch nicht zuverlässig erfasst werden.
- Schülerinnen und Schüler mit Migrationshintergrund müssen über ausreichende Deutschkenntnisse verfügen. Ein schlechtes Testergebnis könnte bei ungenügenden Sprachkenntnissen sowohl auf die Leseleistung als auch auf mangelnden Wortschatz zurückführbar sein. Die Testergebnisse sind in diesem Fall also nicht eindeutig interpretierbar.
- Die Beurteilung von Kindern, die eine Klassenstufe wiederholt haben, sollte generell mit Bedacht erfolgen. Das Lesen stellt eine Leistung dar, die sich kontinuierlich entwickelt und nicht nur durch entsprechende Inhalte des Schulunterrichts gefördert wird. Die Zuordnung solcher Kinder zu einer geeigneten Referenzgruppe ist deshalb nicht ganz unproblematisch. Generell sollten Kinder, die bereits eine Klassenstufe wiederholt haben, eher gemäß der bisherigen Gesamtzahl an Schuljahren anstatt gemäß der aktuell besuchten Klassenstufe eingeordnet werden, insbesondere wenn der Grund für das Wiederholen der Klasse in schlechter Leistung begründet liegt. Im Einzelfall kann allerdings ein Kind auch eine Klassenstufe wiederholen, weil in einem Schuljahr keine ordentliche Beschulung möglich war (z. B. aus Krankheitsgründen). In solchen Fällen kann auch eine Beurteilung in Bezug auf die aktuell besuchte Klassenstufe sinnvoll sein.
- Zur Durchführung des Computertests sollte das Kind grundlegende Erfahrungen am Computer gesammelt haben. Insbesondere sollte das Kind mit der Computermaus umgehen können. Dies muss eventuell vorher nochmals geübt werden. Bei Kindern, die über gar keine entsprechende Erfahrung verfügen, sollte der Test lieber als Papierform appliziert werden. Obwohl der Test am Computer in der Regel ohne das Eingreifen des Testleiters/der Testleiterin vonstattengeht, sollte dieser/diese die Durchführung kontinuierlich überwachen, um im Falle von Anwendungsschwierigkeiten sofort Hilfestellung leisten zu können. Dabei darf das Kind jedoch auf keinen Fall hinsichtlich der Aufgabenbeantwortung beeinflusst werden. Sollten Anwendungsprobleme während der Durchführung eines der Untertests auftreten, so können die Ergebnisse dieses Tests nicht bewertet werden. Der Test muss gegebenenfalls abgebrochen und nach einem angemessenen Zeitraum (ca. 2 Wochen) komplett wiederholt werden.
- Bei Gruppentestungen ist darauf zu achten, dass die Kinder nicht voneinander abschreiben. Dies gilt insbesondere, da im Gegensatz zur Vorgängerversion aus methodischen Gründen keine Pseudoparalleltestversion mehr zur Verfügung steht. Unterhaltungen während der Testdurchführung sind zu unterbinden. Bei großen Gruppen empfiehlt es sich, dass zwei oder mehr Testleiter die Durchführung beaufsichtigen. Verwenden Sie wenn möglich Sichtschutzvorrichtungen zwischen den Testteilnehmern.

8.5 Anleitung für die Testung mit der Papierform

8.5.1 Allgemeine Angaben zur Testdurchführung

Vor der Durchführung des Tests müssen zunächst die persönlichen Daten auf der Vorderseite eingetragen werden. Bei Gruppentestungen kann diese Aufgabe nach dem Austeilen der Testhefte von den älteren Schülern selbst übernommen werden. Bei Schülern der ersten und zweiten Klassenstufe ist es sinnvoll, nur den Vor- und Nachnamen eintragen zu lassen und die restlichen Daten nach der Testung zu ergänzen. Die Angaben zu Schule, Ort, Geburtsdatum, Geschlecht und Muttersprache spielen für die Testauswertung zunächst keine Rolle, können aber für die Interpretation und Gutachtenstellung wichtig oder hilfreich sein.

Anders als bei der Computertestung muss die Bearbeitungszeit bei der Papierform durch den Testleiter/die Testleiterin selbst gestoppt werden. Es ist zwingend notwendig, diese Zeiten genau einzuhalten! Bitte verwenden Sie deshalb eine sekundengenaue Stoppuhr. Eine Übersicht über die Bearbeitungszeiten findet sich in Tabelle 8.1.

Die Kinder erhalten zu Beginn des Tests eine allgemeine Instruktion. Wenn Sie mit einem Kind oder einer Klasse noch nicht oder kaum vertraut sind, dann kann es (vor allem in der Einzeltestung) eventuell notwendig sein, vor dieser kurzen Instruktion ein „aufwärmendes" Gespräch zu führen. Das Ziel dieses Gesprächs besteht darin, Vertrauen zu bilden und dem Kind bzw. den Kindern somit die Angst vor der Testung zu nehmen.

Vor jedem neuen Untertest werden die Kinder mündlich instruiert und die Instruktionsaufgaben werden zusammen bearbeitet. Zur Erleichterung der Testdurchführung steht Ihnen hierfür eine separate Instruktionskarte zur Verfügung.

8.5.2 Allgemeine Instruktion vor Beginn des Tests

Beginnen Sie zunächst mit einer allgemeinen Instruktion:

> *Ich möchte mit dir ein paar Leseaufgaben machen, um herauszufinden, wie gut du schon lesen kannst. Wir werden drei verschiedene Arten von Aufgaben machen.*

Ausfüllen der persönlichen Daten für Schüler der Klassenstufen 1 und 2:

> *Bitte schreibe zunächst deinen Namen auf das Blatt. Schreibe deinen Vornamen in die erste Zeile und deinen Nachnamen in die zweite Zeile.* (Zeigen Sie mit dem Finger auf die entsprechenden Linien des Testheftes. Warten Sie anschließend, bis alle Kinder ihren Namen aufgeschrieben haben.)

Ausfüllen der persönlichen Daten für Schüler der Klassenstufen 3 bis 7:

> *Bitte fülle zunächst die Zeilen hier oben aus.* (Zeigen Sie mit dem Finger auf die Linien ihres Demonstrationstestheftes bis zum Geburtsdatum.) *Beginne mit deinem Vor- und Nachnamen und arbeite bis zu deinem Geburtsdatum.* (Warten Sie, bis alle Zeilen ausgefüllt sind.)
>
> *Jetzt kreuzt du hier an, ob du ein Junge oder ein Mädchen bist.* (Zeigen Sie mit dem Finger auf die entsprechenden Stellen. Warten Sie anschließend, bis alle Kreuze gemacht wurden.)
>
> *Mache ein Kreuz bei ‚deutsch', wenn du zu Hause mit deinen Eltern deutsch sprichst. Mache ein Kreuz bei ‚gemischt', wenn du mit einem Elternteil deutsch und mit dem anderen eine andere Sprache sprichst. Mache ein Kreuz bei ‚andere', wenn du zu Hause mit deinen Eltern nicht deutsch sprichst.* (Zeigen Sie mit dem Finger jeweils auf die entsprechenden Stellen. Warten Sie anschließend, bis alle Kreuze gemacht wurden. Kinder, die aus einem Haushalt stammen, in dem beide Eltern nicht muttersprachlich deutsch sind, sollten auch dann ‚andere' ankreuzen, wenn sie nur ab und zu mit ihren Eltern deutsch sprechen.)

Tabelle 8.1
Zeitlimits für die Testung mit der Papierform

	Wortverständnistest	Satzverständnistest	Textverständnistest
Standardversion	3 Minuten (180 sec.)	3 Minuten (180 sec.)	7 Minuten (420 sec.)
Kurzversion 1–3	3 Minuten (180 sec.)	3 Minuten (180 sec.)	(entfällt)
Kurzversion 4–7	2 Minuten (120 sec.)	2 Minuten (120 sec.)	6 Minuten (360 sec.)