Zweite quantitative Studie

8

Der zweite quantitative Teil der Arbeit basiert auf den in Kapitel 7 extrahierten Faktoren. In Abschnitt 6.1 wurden die Zielvoraussetzungen der Ermittlung des Effektes der Faktoren auf die Schwierigkeit von Testaufgaben für das Instrument zur sprachlichen Variation von Textaufgaben formuliert. Die Ermittlung des Effektes der Faktoren auf die Schwierigkeit von Testaufgaben ist bedeutsam, um Kenntnisse darüber zu erlangen, welche praktischen Implikationen sich für Anpassungsstrategien von mathematischen Textaufgaben ergeben. Das heißt, nur mit der Kenntnis darüber, welcher Faktor welchen Effekt auf die Schwierigkeit einer Testaufgabe hat, kann das Instrument zur sprachlichen Veränderung von mathematischen Textaufgaben effektiv für Anpassungsstrategien von Texten an den Lesenden verwendet werden. Um den Effekt auf die Schwierigkeit einer Testaufgabe zu ermitteln, werden die sprachlichen Faktoren als spezifische Aufgabenmerkmale betrachtet. Eine Methode, den Effekt der sprachlichen Faktoren auf die Aufgabenschwierigkeit festzustellen, ist das LLTM als Erweiterung des Rasch-Modells. Mit dem LLTM wird die Aufgabenschwierigkeit durch die Aufgabenmerkmale berechnet. So ist es möglich, die Berechnung der Aufgabenschwierigkeit zwischen LLTM und Rasch-Modell zu vergleichen und den Effekt der Aufgabenmerkmale auf die Schwierigkeit der Testaufgaben zu bestimmen. Das Ziel dieses Kapitels ist es, durch die beiden geschilderten quantitativen Verfahren die Aufgabenschwierigkeiten zu bestimmen sowie zu vergleichen und den Effekt der Faktoren auf die Aufgabenschwierigkeit zu berechnen.

Überblick: Im zweiten quantitativen Teil der Arbeit wird nur ein Teil der Textaufgaben verwendet, die für die Faktorenanalyse in Kapitel 7 analysiert wurden. Aus diesem Grund ergibt sich für die Zielsetzung und die Teilstichprobe eine daran angelehnte Auswertungsmethode (Abschnitt 8.1). Die Auswertungsgrundlage der Teilstichprobe sind Testaufgaben aus einem längsschnittlichen

D. Bednorz, *Sprachliche Variationen von mathematischen Textaufgaben*,
Bielefelder Schriften zur Didaktik der Mathematik 5,
https://doi.org/10.1007/978-3-658-33003-3_8

Datensatz (Abschnitt 8.1.1). Für die Bestimmung des Effektes der Schwierigkeit ist die Verwendung des Rasch-Modells und des LLTM nötig, wobei Letzteres die Gültigkeit des Rasch-Modells voraussetzt. Dahingehend ergibt sich ein Ablaufmodell für die Analyse (Abschnitt 8.1.2). Die erste Methode zur Berechnung der Aufgabenschwierigkeit ist das Rasch-Modell (Abschnitt 8.2). Dieses ist das bekannteste Verfahren der Item-Response-Theorie (IRT), die insbesondere zur Skalierung und Analyse von Testitems verwendet wird und deren methodische Grundlage vor Verwendung zu klären und auf das Rasch-Modell zu spezifizieren ist (Abschnitt 8.2.1). Zur Testung der Gültigkeit der Modellpassung des Rasch-Modells können inferenzstatistische oder grafische Überprüfungen genutzt werden (Abschnitt 8.2.2). Die Bestimmung der Aufgabenschwierigkeiten des Teildatensatzes führt zu einer Skalierung der Textaufgaben durch das Rasch-Modell (Abschnitt 8.2.3). Dieses Modell wird anschließend durch das LLTM erweitert (Abschnitt 8.3). Zunächst wird dahingehend die methodische Grundlage des LLTM geschildert und der Zusammenhang zum Rasch-Modell geklärt (Abschnitt 8.3.1). Als erstes Ergebnis der Ermittlung der Aufgabenschwierigkeiten durch das LLTM können die ermittelten Aufgabenschwierigkeiten des Rasch-Modells und des LLTM miteinander verglichen werden und die Güte des LLTM kann bestimmt werden (Abschnitt 8.3.2). Als zweites Ergebnis lassen sich die Effekte der Faktoren auf die Aufgabenschwierigkeit bestimmen (Abschnitt 8.3.3). Als drittes lässt sich im Hinblick auf die Passung der bestimmten Aufgabenschwierigkeit des LLTM im Vergleich zum Rasch-Modell prüfen, für welche Testaufgaben eine genaue bzw. weniger genaue Passung der Aufgabenschwierigkeit erreicht wird (Abschnitt 8.3.4). Anhand der Analyse der Passung lassen sich die Erklärungsleistungen der Aufgabenschwierigkeit durch die Verwendung von weiteren quantitativen Methoden (Clusteranalyse, Regressionsanalyse) ausdifferenzieren (Abschnitt 8.3.4). Abschließend werden die Ergebnisse des zweiten quantitativen Teils dieser Arbeit diskutiert (Abschnitt 8.4)

8.1 Auswertungsmethode

Der zweite quantitative Teil dieser Studie erfordert aufgrund der eigenen Zielsetzungen eine darauf ausgelegte Auswertungsmethode.

Überblick (Abschnitt 8.1): Für die Analyse wird nur ein Teil der Gesamtstichprobe einbezogen und es werden zwei IRT-Modelle verwendet, die zunächst geschildert werden. Außerdem erfolgt der Hinweis auf die verwendete Software für die Analyse (Abschnitt 8.1.1). Anschließend wird der Ablauf der Analyse erläutert. Dies

betrifft die Durchführung des Rasch-Modells und des darauf aufbauenden LLTM (Abschnitt 8.1.2).

8.1.1 Auswertungsgrundlage

Grundlage der Auswertung für die Rasch-Analyse sind die Daten aus der längsschnittlichen Studie des Projektes zur Analyse der Leistungsentwicklung in Mathematik (PALMA) (Pekrun et al., 2006; vom Hofe et al., 2002). PALMA umfasst insgesamt sechs Messzeitpunkte, in deren Verlauf die Klassen 5–10 getestet worden sind (vom Hofe et al., 2002). Wie für lang angesetzte längsschnittliche Studien üblich, variiert N für die einzelnen Messzeitpunkte deutlich (N_{MZP1} = 2070, N_{MZP2} = 2070, N_{MZP3} = 2395, N_{MZP4} = 2409, N_{MZP5} = 2521, N_{MZP6} = 1943). Es stammen von der Gesamtstichprobe 68 Aufgaben aus dem Itemsatz von PALMA (MZP_1 = 10 Items, MZP_2 = 8 Items, MZP_3 = 17 Items, MZP_4 = 10 Items, MZP_5 = 15 Items, MZP_6 = 13 Items). Von den vorhandenen Items wurden 16 als Ankeritems für mehrere Messzeitpunkte verwendet. So ergeben sich insgesamt N_{Items} = 47 Items, wie in Abschnitt 6.4 genannt wird, die zur Skalierung in das Rasch-Modell eingebracht werden konnten.

8.1.2 Ablauf der Analyse

Für die zweite quantitative Analyse wurden geeignete Testaufgaben aus der PALMA-Studie selektiert. Grundlage zur Selektion waren Items zur inner- und außermathematischen Modellierung mit offenem Aufgabenformat. Für die weitere Analyse der Bestimmung der Aufgabenschwierigkeiten durch das LLTM war es notwendig, die Aufgaben durch das Rasch-Modell zu skalieren. Das Rasch-Modell muss gelten, damit eine Erweiterung durch das LLTM erfolgen kann. Aus diesem Grund erfolgte die Skalierung der 47 Testaufgaben zunächst durch das Rasch-Modell. Aufgrund des Ankerdesigns und der längsschnittlichen Untersuchungsanlage konnte zur Prüfung des Rasch-Modells nicht der Andersens LR-Test mit dem Split-Kriterium des Mittelwerts verwendet werden. Aufgrund von datenschutzrechtlichen Aspekten konnten ebenfalls keine personenbezogenen Informationen zur Aufteilung des Datensatzes genutzt werden. Entsprechend verblieb zur Prüfung der Passung des Rasch-Modells die Ermittlung von Infit- und Outfit-Statistiken sowie die grafische Überprüfung durch den ICC-Plot und eine Personen-Item-Darstellung bzw. Wright-Darstellung als grafische Option

zur Prüfung des Rasch-Modells. Zur vollständigen Darstellung der unterschiedlichen Möglichkeiten der Überprüfung der Geltung des Rasch-Modells werden in Abschnitt 8.2.2 sowohl die inferenzstatistischen Möglichkeiten als auch die grafischen Optionen zur Prüfung des Modells erörtert.

Für das Rasch-Modell wurden die Personenparameter θ und die Itemparameter β und für das LLTM die Itemparameter β und der schwierigkeitsgenerierende Effekt η berechnet. Diese Berechnung erfolgte durch das R-Paket *eRm* (Mair & Hatzinger, 2007). Für das anschließende LLTM wurden die z-standardisierten Regressionswerte der Faktoren über den Median dichotomisiert. Damit entspricht eine 1 einer hohen Ausprägung auf einen Faktor. Für diese Aufgaben kann der jeweilige sprachliche Faktor als bedeutend interpretiert werden. Eine 0 entspricht einer negativen Ausprägung auf einen Faktor. Der sprachliche Faktor kann entsprechend dahingehend gedeutet werden, dass er keine (kaum) Bedeutung für die jeweilige Testaufgabe hat. Die Faktoren werden für das LLTM als Aufgabenmerkmale (kognitive Operatoren) interpretiert, die zur Lösung der Aufgaben benötigt werden.

Durch die in Kapitel 5 erläuterten theoretischen und empirischen Befunde lassen sich für Textmerkmale Ableitungen bezüglich der Erwartungen des Effektes auf die Aufgabenschwierigkeit treffen. Verdeutlicht werden kann dies, wenn die Ergebnisse der Faktorenanalyse, die in Abbildung 8.1 zusammengefasst sind, betrachtet werden. Die in Abbildung 8.1 dargestellte Systematisierung der Textmerkmale erfolgte nicht nach dem Effekt auf die Schwierigkeit des Textes, sondern nach dem gemeinsamen Vorkommen. Das bedeutet, dass je Faktor Textmerkmale vorkommen können, die als schwierigkeitsgenerierend betrachtet werden können oder einen gegenteiligen Effekt auf die Textschwierigkeit haben können.

Es wird nach Kapitel 5 angenommen, dass die Textschwierigkeit einen Einfluss auf die Aufgabenschwierigkeit besitzt. Exemplarisch ist die geringe Erwartbarkeit des Effektes auf die Schwierigkeit des Textes insbesondere für den komprimierenden, beschreibenden und informativen Faktor erkennbar. Die jeweiligen Faktoren sind durch Textmerkmale bestimmt, die auf der einen Seite einen positiven Effekt und auf der anderen Seite einen negativen Effekt auf die Textschwierigkeiten aufweisen können. Beispielsweise hat der komprimierende Faktor zum einen das Textmerkmal *Passiv*, das erwartungsgemäß für einen positiven Effekt auf die Textschwierigkeit gelten kann, und zum anderen Text-Bild-Referenzen, für die tendenziell angenommen werden kann, dass sie einen negativen Effekt auf die Textschwierigkeit aufweisen (vgl. Abschnitt 5.2.4). Entsprechend ergeben sich keine spezifischen Erwartungen für den schwierigkeitsgenerierenden Effekt für

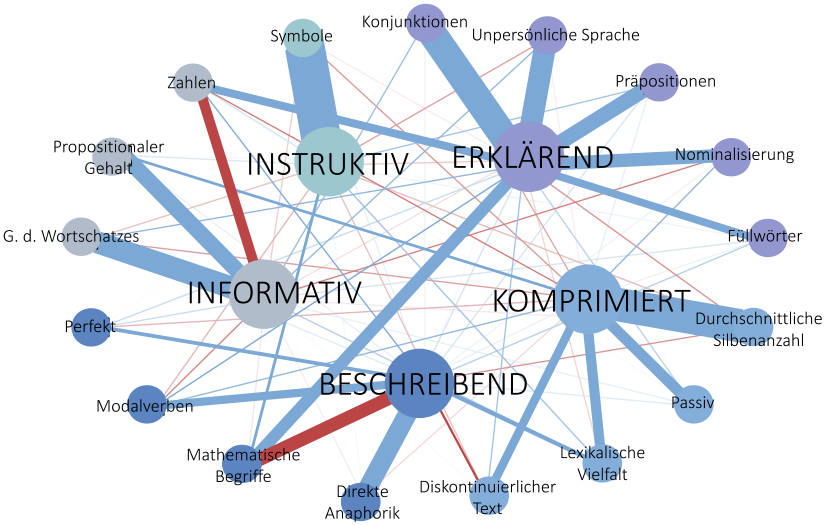

Abbildung 8.1 Zusammenfassung der Faktorenanalyse nach Interpretation der Faktoren. (Eigene Erstellung)

diese Faktoren, stattdessen muss der Effekt auf die Schwierigkeit im Hinblick auf die Ergebnisse der Datenanalyse interpretiert werden.

8.2 Rasch-Modell

Überblick (Abschnitt 8.2): Das Rasch-Modell ist ein Verfahren aus der IRT. Diese wird insbesondere zur Skalierung von Testitems verwendet. Das Rasch-Modell wird genutzt, um die Aufgabenschwierigkeiten für den Teildatensatz in dieser Arbeit zu berechnen. Zur Darstellung der Berechnungsgrundlage wird zunächst die methodische Grundlage der IRT dargestellt und für das Rasch-Modell konkretisiert (Abschnitt 8.2.1). Zur Prüfung der Modellvoraussetzungen existieren inferenzstatistische und grafische Verfahren, die im Überblick dargestellt werden sollen (Abschnitt 8.2.2). Zum Schluss erfolgt die Ergebnisdarstellung der

Skalierung der $N = 47$ ausgewählten Testaufgaben durch das Rasch-Modell (Abschnitt 8.2.3).

8.2.1 Methodische Grundlagen der Item-Response-Theorie

Gemäß Geiser und Eid (2010) werden IRT Modelle zur Analyse und Skalierung von Test- und Fragebogenitems verwendet. Die Modellierungsgrundlage für IRT-Modelle ist die Beziehung zwischen Probandinnen und Probanden und deren Antwortverhalten auf eine Testaufgabe in Form einer Frage oder Feststellung. In Bezug auf das erfolgreiche Absolvieren der Testaufgabe, die den zentralen Gegenstand der IRT darstellt (im Vergleich zur klassischen Testtheorie, bei der der Test betrachtet wird), kann bei einer gewissen Stichprobengröße von Probandinnen und Probanden sowie Testaufgaben der Schwierigkeitsgrad der Items abgeschätzt und die Erfolgswahrscheinlichkeiten für die Testaufgaben können modelliert werden (Geiser & Eid, 2010; Kean & Reilly, 2014; Kleine, 2004).

Nach Geiser und Eid (2010) ist für alle IRT-Modelle die Annahme identisch, dass den beobachtbaren Antwortverhalten (manifeste Variable) eine nicht beobachtbare Eigenschaft (latente Variable) zugrunde liegt. Es besteht ein wahrscheinlichkeitsfunktionaler Zusammenhang zwischen latenter und manifester Variablen. Damit wird die Wahrscheinlichkeit ausgedrückt, dass ein bestimmtes Antwortverhalten zustande kommt, in Beziehung zu der Eigenschaft, die bei der Probandin oder beim Probanden erhoben wird (Personenparameter), und Spezifika der Items (Itemparameter). Aus diesem Grund werden IRT-Modelle unter dem Begriff der *probabilistischen Testtheorie* zusammengefasst.

Grundlegend für die IRT sind die Itemkennlinien bzw. Itemcharakteristika, von denen alle anderen Konstrukte der Theorie abhängen (Baker, 2001; Kleine, 2004). In Abbildung 8.2 ist beispielhaft der ogivenförmige Funktionsverlauf der logistischen Funktion dargestellt, die mit $f(x) = \frac{e^x}{1+e^x}$ definiert ist und als *Item Charakteristic Curve* (ICC) bezeichnet wird (Kleine, 2004).

Die Itemkennlinien besitzen zwei zu interpretierende Eigenschaften. Zum einen kann über die latente Dimension (z. B. Fähigkeit), im Koordinatensystem auf der Abszisse abgetragen, die Schwierigkeit des Items in Bezug auf Personenfähigkeiten begutachtet werden. So kann für die Items jeweils die Wahrscheinlichkeit der Lösung bei einer bestimmten latenten Dimension betrachtet werden (Kleine, 2004). Bei geringen Fähigkeiten erhalten nur leichte Items eine gewisse Wahrscheinlichkeit der Lösung. Demgegenüber ist ein höherer Fähigkeitswert notwendig, um eine gewisse Lösungswahrscheinlichkeit bei schwierigen Items zu erhalten. Zum anderen haben die Itemkennlinien die Eigenschaft der

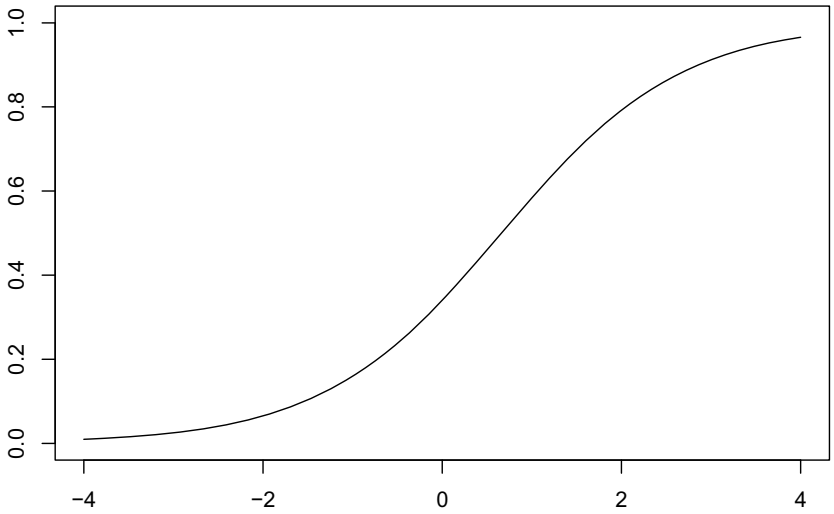

Abbildung 8.2 Beispiel des ogivenförmigen Funktionsverlaufs der logistischen Funktion. (Eigene Erstellung)

Diskrimination bzw. Trennschärfe, mit der geprüft werden kann, wie gut sich Items unterscheiden lassen (Baker, 2001; Geiser & Eid, 2010; D. Rasch et al., 2011). Deutlich wird die Diskrimination durch die Steilheit der Itemkennlinie in ihrem mittleren Abschnitt. Je steiler die Kurve, desto genauer, und je flacher die Kurve, desto ungenauer kann das Item diskriminieren. Das bedeutet, dass die Wahrscheinlichkeit einer korrekten Antwort bei niedriger Fähigkeit annähernd die gleiche ist wie bei einer hohen Fähigkeit.

Wenn bei einem IRT-Modell angenommen wird, dass alle Items dieselbe latente Dimension mit unterschiedlichen Itemschwierigkeiten bei gleicher Trennschärfe besitzen, wird von einem *1-parametrigen Rasch-Modell* gesprochen (da nur eine latente Dimension existiert) (Geiser & Eid, 2010). Das Rasch-Modell wird als das unkomplizierteste IRT-Modell bezeichnet, da es die Wahrscheinlichkeit einer Aufgabenlösung in Abhängigkeit von der Differenz zwischen Fähigkeit und Itemschwierigkeit modelliert (Geiser & Eid, 2010; Moosbrugger, 2012; G. Rasch, 1960).

Bei dem Rasch-Modell wird angenommen, dass die Lösungswahrscheinlichkeit von der Schwierigkeit des Items β_j und von der Personenfähigkeit θ_i abhängt. Die Modellgleichung für das Rasch-Modell lautet (Bühner, 2011; Eid & Schmidt,

2014; Moosbrugger, 2012; G. Rasch, 1960; Strobl, 2015):

$$P(U_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$$

oder umformuliert:

$$P(U_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

Die Wahrscheinlichkeit zur Lösung ist damit eine bedingte Wahrscheinlichkeit in Abhängigkeit von der Itemschwierigkeit und Personenfähigkeit (Kleine, 2004; D. Rasch et al., 2011; Strobl, 2015). Bei einer Betrachtung der Differenz von θ_i und β_j, wird deutlich, dass die Lösungswahrscheinlichkeit davon abhängt, wie fähig eine Person im Vergleich zu der Schwierigkeit einer Aufgabe ist (Kleine, 2004; Strobl, 2015). Wenn einerseits die Person eine höhere Fähigkeit als die Schwierigkeit der Aufgabe erfordert besitzt ($\theta_i > \beta_j$), ergibt sich eine positive Differenz und die Lösungswahrscheinlichkeit ist groß (Strobl, 2015). Wenn andererseits die Person eine niedrigere Fähigkeit als die Schwierigkeit der Aufgabe erfordert besitzt ($\theta_i < \beta_j$), ergibt sich eine negative Differenz und die Lösungswahrscheinlichkeit ist niedrig (Strobl, 2015).

Die in der Gleichung genannte bedingte Wahrscheinlichkeit wird durch den ICC-Plot in Abbildung 8.2 grafisch dargestellt. Nach der Modellgleichung ergibt sich zwischen Itemschwierigkeit und Personenfähigkeit ein logistischer Zusammenhang (Geiser & Eid, 2010; Kleine, 2004).

Das Rasch-Modell besitzt zwei bedeutsame statistische Grundannahmen. Die erste ist die der suffizienten Statistiken. Die suffizienten Statistiken ergeben sich beim Rasch-Modell aus den Zeilenrand-Summen $r_i = \sum_{j=1}^{m} u_{ij}$ und den Spalten-Randsummen $s_j = \sum_{i=1}^{n} u_{ij}$ (Geiser & Eid, 2010; Strobl, 2015). Suffiziente Statistik bedeutet dahingehend, dass es zur Schätzung der Fähigkeit einer Person irrelevant ist, welche Aufgaben sie gelöst hat, und dass nur von Bedeutung ist, wie viele Aufgaben von der Person in Summe gelöst wurden. Ebenfalls ist für die Abschätzung der Itemschwierigkeit unerheblich, welche Person welche Aufgabe gelöst hat. Hierbei ist die Summe der gesamten Probandinnen und Probanden für die Aufgabe entscheidend.

Die zweite statistische Grundannahme beim Rasch-Modell ist die der lokalen stochastischen Unabhängigkeit (Geiser & Eid, 2010; D. Rasch et al., 2011; Strobl, 2015). Für mehrere Aufgaben im Rasch-Modell muss entsprechend gelten, dass die Lösungswahrscheinlichkeit einer Aufgabe unabhängig davon ist, ob

eine andere Aufgabe gelöst werden konnte (Geiser & Eid, 2010; Strobl, 2015). Daraus ergibt sich für die Itemkonstruktion, dass die Lösungswahrscheinlichkeit bei Teilaufgaben nicht von Teillösungsschritten abhängt.

Ein weiteres Kriterium des Rasch-Modells ist die spezifische Objektivität. Das bedeutet zum einen, dass es zur Prüfung der Fähigkeit von Personen unerheblich ist, mit welcher Lösungswahrscheinlichkeit einer konkreten Aufgabe die Personen verglichen werden. Zum anderen ist ein Vergleich der Schwierigkeit der Items unabhängig von den Fähigkeiten möglich (Geiser & Eid, 2010; D. Rasch et al., 2011; Strobl, 2015).

8.2.2 Prüfen der Modellpassung

Zur Prüfung der Modellpassung des Rasch-Modells ergeben sich unterschiedliche Prüfverfahren. Sie werden in inferenzstatistische Modelltests (z. B. Andersen-Test) und nicht inferenzstatistische Modelltests (grafische Modelltests) differenziert.

Eine Möglichkeit zur inferenzstatistischen Prüfung des Modells ist der *Likelihood-Ratio Test* (LR-Test) nach Andersen (1973). Für den LR-Test wird die Stichprobe geteilt. Das Teilungskriterium kann beispielsweise der Mittelwert oder Median der Lösungshäufigkeit oder aus Personenausprägungen wie dem Geschlecht bestehen (Bühner, 2011). Durch den LR-Test wird ein empirischer χ^2-Wert berechnet und mit dem kritischen χ^2-Wert mittels der Anzahl der vorhandenen Freiheitsgrade verglichen (Bühner, 2011; D. Rasch et al., 2011). Ein signifikantes Ergebnis des LR-Test bedeutet, dass die Annahme verworfen werden muss, dass das Rasch-Modell gilt. Ein nicht signifikantes Ergebnis des LR-Tests verweist darauf, dass das Rasch-Modell gilt und beibehalten werden kann (Andersen, 1973). Neben dem LR-Test basiert als weiterer inferenzstatistischer Modelltest der Wald-Test auf dem Vergleich von Gruppen. Für den Wald-Test wird jedoch jedes einzelne Item betrachtet. Dies kann dazu beitragen, unpassende Items aus dem Modell zu eliminieren (D. Rasch et al., 2011). Wie in Abschnitt 8.1.2 benannt, kann aufgrund des Ankerdesigns und fehlender personenbezogener Angaben, weder der LR-Test nach Andersen noch der Wald-Test zur Prüfung des Rasch-Modells verwendet werden.

Neben dem LR-Test bietet sich noch die Prüfung über Infit- und Outfit-Statistiken an. Hierbei werden nicht Gruppen verglichen, sondern die quadrierten Residuen (Ayala, 2009). Die Interpretation der Infit- und Outfit-Mean-Square (MSQ) erfolgt durch Ermittlung der Passung der Itemschwierigkeit und der Personenparameter. Nach Wright und Linacre (1994) sind Werte, die einen MSQ über

2 erreichen, als nicht akzeptabel anzusehen. Solche MSQ, die kleiner als 1.5–2.0 sind, können aufgrund der nicht zu verschlechternden Statistik beibehalten werden. Als optimal bewertet werden MSQ zwischen 0.5 und 1.5. Ein Infit von 1 weist auf perfekte Passung zum Modell hin. Items, die kleiner sind als 1, stellen einen Overfit dar und diskriminieren tendenziell stark.

Eine nicht inferenzstatistische Modelltestung kann über einen Goodness-Of-Fit(GOF)-Plot erfolgen. Die Voraussetzung für einen GOF-Plot ist die vorherige Berechnung durch den Andersen LR-Test. Damit stellt die Darstellung über einen GOF-Plot die Visualisierung des Ergebnisses der LR-Tests dar. Der GOF kann dahingehend interpretiert werden, dass die Passung eines Items umso geringer ist, je weiter es von der Diagonalen entfernt ist (Andersen, 1973). Außerdem können zur genauerer Interpretation Konfidenzintervall-Ellipsen dargestellt werden, um zu prüfen, ob ein Item eine hinreichende Passung erreicht.

Daneben bietet sich zur grafischen Modellpassung die Prüfung durch den ICC-Plot, einen Personen-Item-Plot oder die Wright-Map an. Mit dem ICC-Plot kann die Funktionsweise der Items charakterisiert und die theoretisch angenommene Verteilung empirisch überprüft werden. Im ICC-Plot können unübliche Verläufe von Items für die Passung zum Rasch-Modell registriert und Items daraufhin selektiert werden. Eine genauere Überprüfung ermöglichen der Personen-Item-Plot und die Wright-Map, die beide die Ausprägungen der Items und der Personenparameter auf der latenten Dimension darstellen. Dahingehend kann durch die grafische Überprüfung abgebildet werden, ob durch die vorhandenen Items die Fähigkeiten der Personen in ausreichendem Maße gemessen werden können.

8.2.3 Ergebnisse Rasch-Modell

Zur Darstellung der Ergebnisse der Skalierung der mathematischen Testaufgaben werden deskriptive Statistiken, Statistiken zur Reliabilität und der geschätzten β-Werte zur Schwierigkeit der einzelnen Testaufgaben berichtet.

Nach dem in Abschnitt 8.1.2 erläuterten Vorgehen wurden $N = 47$ mathematische Testaufgaben der unterschiedlichen Messzeitpunkte aus PALMA in einen zusammengeführten Datensatz übertragen, um darauf aufbauend die Re-Analyse der ausgewählten Textaufgaben mit offenem Aufgabenformat für das Rasch-Modell durchzuführen. Für das Rasch-Modell ergibt sich eine Conditional Log Likelihood von -26765.2 mit einer Parameteranzahl von $\mathrm{N} = 46$. Die Varianz der Aufgabenschwierigkeit liegt bei 2.31 ($SD = 1.52$). Der kleinste Wert für die Aufgabenschwierigkeit beträgt $\beta_{\min} = -3.56$. Der größte erreichte Wert

für die Aufgabenschwierigkeit liegt bei $\beta_{\max} = 3.44$. Die Messgenauigkeit für die vorliegenden Testaufgaben liegt bei $\text{EAP}_{\text{Reliabilität}} = 0.613$ ($\text{WLE}_{\text{Reliabilität}} = 0.468$). Aufgrund des geschilderten Ankerdesigns haben nicht alle Schülerinnen und Schüler alle Aufgaben bearbeitet; entsprechend ist nicht von einer hohen Reliabilität auszugehen. Aus diesem Grund ist die berichtete Reliabilität als zufriedenstellend anzusehen.

In Abbildung 8.3 sind für die einzelnen Testaufgaben die charakteristischen Itemkennlinien dargestellt. Wie in Abbildung 8.2 exemplarisch dargestellt, entspricht der empirisch ermittelte Verlauf der 47 Items aus den Aufgabensamples dem theoretisch angenommenen ogivenförmigen Funktionsverlauf der logistischen Funktion, der durch das Rasch-Modell postuliert wird und in Abschnitt 8.2.1 erläutert wurde. Das leichteste Item *Hundefutter* (Nummer 3, da unsortiert, vgl. Tabelle 8.2) erreicht bei einer Traitausprägung der latenten Dimension von 0 eine fast hundertprozentige Lösungswahrscheinlichkeit. Im Gegensatz dazu erreicht das schwerste Item *Zahlenrätsel* (Nummer 45, da unsortiert, vgl. Tabelle 8.1) nur eine Lösungswahrscheinlichkeit, die knapp über 0 ist. Der ICC-Plot in Abbildung 8.3 macht deutlich, dass sich die Traitausprägung der

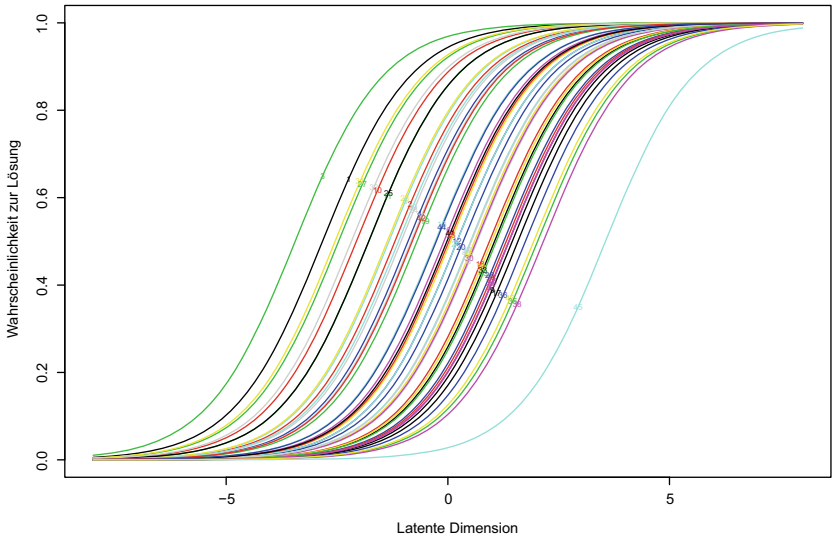

Abbildung 8.3 Charakteristische Itemkennlinien (ICC) der Testaufgaben nach Skalierung durch das Rasch-Modell. (Eigene Erstellung)

Testaufgaben breit auf der latenten Dimension erstreckt und so unterschiedliche Fähigkeiten gemessen werden können.

In Tabelle 8.1 (für negative β) und fortgesetzt in Tabelle 8.2 (für positive β) sind die Ergebnisse der Skalierung der Aufgabenschwierigkeit der Testaufgaben dargestellt. Für die Auflistung in den Tabellen 8.1 und 8.2 wurden die geschätzten Aufgabenschwierigkeiten von aufsteigender nach absteigender Schwierigkeit sortiert. Die Testaufgabe *Brinkmeier* wurde in allen Messzeitpunkten als Item verwendet. Die restlichen Testaufgaben wurden jeweils nur für einen, zwei, drei oder vier Messzeitpunkte genutzt.

Tabelle 8.1 Schätzung der Aufgabenschwierigkeiten des Rasch-Modells (negatives β)

Item	Geschätztes β	Standardfehler	Infit-MSQ	Outfit-MSQ
Zahlenrätsel	−3.56	0.14	0,76	0,60
Diagonale	−2.15	0.07	0,81	0,62
Grünberg	−2.02	0.07	0,77	0,57
Litfaßsäule	−1.94	0.10	0,89	0,86
Harry Potter	−1.75	0.07	1,04	1,00
Pause A	−1.58	0.07	0,98	0,94
Frau Amann	−1.45	0.10	0,76	0,65
Zugspitzbahn	−1.38	0.07	0,81	0,72
Begründung	−1.38	0.13	0,87	0,78
Flussbreite	−1.37	0.10	0,98	0,94
Leiter	−1.32	0.08	0,79	0,73
Fernseher	−1.26	0.13	0,86	0,58
Pause B	−1.09	0.09	1,15	1,24
Schnellfahrer	−1.05	0.10	0,84	0,71
Brot	−0.99	0.08	0,96	0,90
Darlehen	−0.94	0.07	1,06	1,01
Rechteck	−0.66	0.06	0,83	0,71
Füße	−0.64	0.11	0,96	1,10
Burgturm	−0.60	0.07	0,98	0,94

(Fortsetzung)

Tabelle 8.1 (Fortsetzung)

Item	Geschätztes β	Standardfehler	Infit-MSQ	Outfit-MSQ
Goethe B	−0.53	0.08	0,92	0,80
Freibad	−0.51	0.08	0,88	0,82
Farbe	−0.35	0.11	0,91	0,84
Rechenausdruck	−0.21	0.07	1,09	1,03
Lernpro 3	−0.21	0.12	0,80	0,77
Glühlampe	−0.21	0.12	0,81	0,69
Seehund	−0.02	0.09	0,94	0,90

Tabelle 8.2 Schätzung der Aufgabenschwierigkeiten des Rasch-Modells (positives β)

Item	Geschätztes β	Standardfehler	Infit-MSQ	Outfit-MSQ
IPod	0.02	0.08	0,96	0,95
Schulfahrt	0.09	0.08	0,77	0,64
Flakes	0.27	0.11	0,90	0,87
Lena	0.29	0.06	1,00	1,07
Quark	0.70	0.07	0,86	0,78
Bratkartoffeln	0.80	0.08	0,89	0,82
Arbeitsamt	0.83	0.07	0,95	1,01
Lotto	0.90	0.06	0,88	0,84
Aktion	1.05	0.06	0,89	0,86
Stausee	1.11	0.08	0,94	0,89
Schoko	1.19	0.06	0.83	0.77
Schulfest	1.33	0.07	1,01	1,03
Erbsen	1.36	0.08	1,00	0,96
Füssen	1.78	0.10	0,70	0,60
Dreieckswinkel	1.79	0.08	0,92	0,87
Kontostand A	2.06	0.06	0,93	0,90
Kontostand B	2.18	0.10	1,03	1,17
Andreas	2.46	0.08	1,05	1,12
Kinderzimmer	2.54	0.08	1,09	1,19

(Fortsetzung)

Tabelle 8.2 (Fortsetzung)

Item	Geschätztes β	Standardfehler	Infit-MSQ	Outfit-MSQ
Brinkmeier	2.83	0.05	0.98	1.22
Hundefutter	3.44	0.07	0.82	0.89

Die anspruchsvollste Testaufgabe nach den Schätzungen des Rasch-Modells ist die Testaufgabe *Zahlenrätsel* aus dem sechsten Messzeitpunkt (vgl. Tabelle 8.1 und Tabelle 8.3). Die leichteste Testaufgabe nach der Schätzung durch das Rasch-Modell ist die Aufgabe *Hundefutter* aus dem ersten und zweiten Messzeitpunkt (vgl. Tabelle 8.2 und Tabelle 8.3).

Inwieweit die Testaufgaben die Fähigkeiten der Personen in genauem Maße messen, wird durch die Personen-Item-Darstellung in Abbildung 8.4 deutlich. In der Abbildung wird im oberen Bereich die Verteilung der Fähigkeitsparameter der Testpersonen dargestellt. Im unteren Bereich werden die Testaufgaben sortiert nach der Aufgabenschwierigkeit an der latenten Dimension angeordnet.

Durch die in Abbildung 8.4 dargestellten Personen-Item-Darstellungen ist es möglich, grafisch zu prüfen, ob die verwendeten Testaufgaben die Traitausprägungen der Personenfähigkeiten auf der latenten Dimension messen können. In Abbildung 8.4 ist zu erkennen, dass der größte Anteil der Verteilung der Personenfähigkeiten von den eingesetzten Testaufgaben gemessen wird. Die grafische Überprüfung bietet jedoch nur einen ungenauen Hinweis darauf, inwieweit eine hohe Passung der Testaufgaben an das Rasch-Modell vorliegt. Aus diesem Grund werden die Infit- und Outfit-Statistiken betrachtet. Neben der grafischen Überprüfung weisen die Infit- und Outfit-Statistiken der skalierten Testaufgaben in den Tabellen 8.1 und 8.2 auf die hohe Passung der Testaufgaben zum Rasch-Modell hin. Die Infit-Werte liegen in einem Bereich von 0.70–1.15, die Outfit-Werte in einem Bereich von 0.57–1.24. Die Werte liegen in dem von Wright und Linacre (1994) definierten Bereich zwischen 0.5 und 1.5, der für Testaufgaben für die Messung als produktiv erachtet werden kann.

Tabelle 8.3 Darstellung der leichtesten und schwierigsten Testaufgaben in dem vorliegenden Datensatz

Aufgabe Hundefutter: Kathrin hat fünf Dosen Hundefutter gekauft. Zusammen kosten die Dosen 10,50 €. Wie viel kostet eine Dose?	Aufgabe Zahlenrätsel: Die Differenz zweier Zahlen beträgt 7. Multipliziert man die kleinere Zahl mit 2 und die größere mit 3, so beträgt die Differenz 25. Wie lauten die beiden Zahlen?

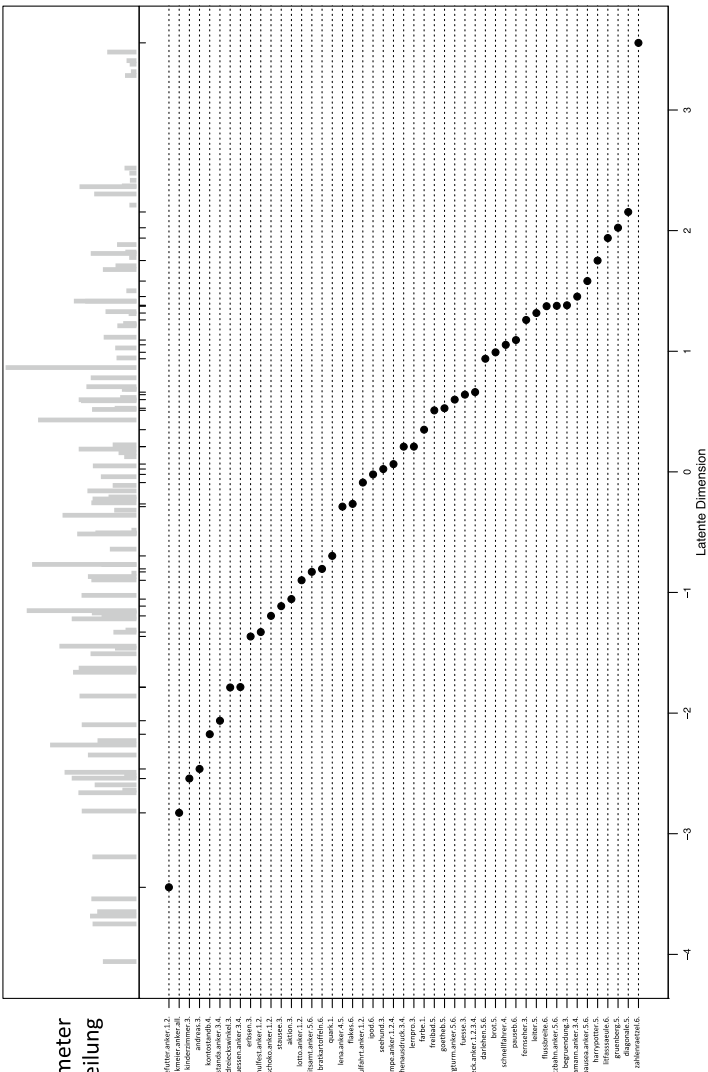

Abbildung 8.4 Personen-Item-Darstellung der Ausprägungen der Testaufgaben und der Personenfähigkeiten auf der latenten Dimension. (Eigene Erstellung)

Die Infit- und Outfit-Statistiken weisen, neben der grafischen Prüfung, auf eine hohe Passung der Testaufgaben zum Rasch-Modell hin. Aus diesem Grund wird das Rasch-Modell für die Skalierung der Testaufgaben angenommen und die Nullhypothese, dass das Rasch-Modell nicht gilt, wird verworfen. Aufgrund der Geltung des Rasch-Modells kann die Erweiterung durch das LLTM erfolgen.

8.3 Linear-logistisches Testmodell

Aufgrund der in Abschnitt 8.2.3 erläuterte Annahme, dass das Rasch-Modell gilt, kann die Erweiterung durch das LLTM erfolgen. Durch das LLTM werden die im Rasch-Modell verwendeten 47 Aufgabenschwierigkeiten durch die 5 dichotomisierten Faktorenwerte der sprachlichen Faktoren ersetzt (vgl. Abschnitt 8.1.2).

Überblick (Abschnitt 8.3): Das LLTM ist eine Möglichkeit, die durch das Rasch-Modell erstellten Aufgabenschwierigkeiten durch Aufgabenmerkmale der Testaufgaben zu schätzen. Das LLTM dient in dieser Arbeit zur Schätzung der durch das Rasch-Modell bestimmten Aufgabenschwierigkeiten, deren methodische Grundlage vorgestellt wird (Abschnitt 8.3.1). Aufgrund der methodischen Anlage des LLTM ist es möglich, die Aufgabenschwierigkeiten des Rasch-Modells mit den geschätzten Aufgabenschwierigkeiten des LLTM zu vergleichen und so die Passung zwischen beiden Modellen zu evaluieren (Abschnitt 8.3.2). Wenn die Passung der geschätzten Aufgabenschwierigkeiten des LLTM ausreichend ist, kann der Effekt der Faktoren (als Aufgabenmerkmale) bestimmt werden (Abschnitt 8.3.3). Der Vorteil der Verwendung des LLTM ist die Berechnung von geschätzten Aufgabenschwierigkeiten je Testaufgabe. So kann ermittelt werden, für welche Testaufgaben die Schätzung durch die Faktoren genau oder weniger genau erfolgte (Abschnitt 8.3.4). Durch die Ergebnisse dieser differenzierten Betrachtung können die Testaufgaben mithilfe einer Clusteranalyse gruppiert werden und die Erklärungsleistung der Aufgabenschwierigkeiten kann jeweils durch ein Regressionsmodell separiert bestimmt werden (Abschnitt 8.3.5).

8.3.1 Methodische Grundlagen des linear-logistischen Testmodells

Wie in Abschnitt 8.2.1 erläutert, beschreibt das Rasch-Modell sowohl die Personen- als auch die Aufgabenseite bei der Datenanalyse. Es dient damit zur

Vorhersage einer Wahrscheinlichkeitsaussage über die Aufgabenlösung bei einer bestimmten Fähigkeit.

Nach Zimmermann (2016) kann die Aufgabenschwierigkeit des Rasch-Modells als abhängige Variable für weiterführende Analysen verwendet werden – beispielsweise, um den Einfluss bestimmter Aufgabenmerkmale als unabhängige Variablen zu bestimmen. Zur Erklärung der Aufgabenschwierigkeit durch Aufgabenmerkmale ergeben sich zwei häufig genutzte Verfahren. Erstens können über ein Regressionsmodell die Aufgabenschwierigkeit als abhängige Variable, die erklärt werden soll, verwendet werden und über sprachliche Merkmale als unabhängige Variablen erklärt werden. Zweitens existieren neben dem Rasch-Modell weitere IRT-Modelle, die die Aufgabenschwierigkeit nur auf der Seite der Aufgabenmerkmale erklären. Ein solches Modell ist das LLTM, für das die Geltung des Rasch-Modells vorausgesetzt wird (Wilson & Boeck, 2004; Wilson & Moore, 2011). Analysen zeigen die Tendenz, dass eine Erklärung der Aufgabenschwierigkeit durch Regressionsmodelle zu ähnlichen Ergebnissen wie das LLTM führt (Hartig, 2007; Isaac & Hochweber, 2011; Zimmermann, 2016).

Das LLTM wird in der vorliegenden Arbeit verwendet, da es bezüglich der Konzeptualisierung eines Instruments zur sprachlichen Variation von Textaufgaben und der mathematikdidaktischen Betrachtung der Textaufgaben einen besonderen Vorteil aufweist. Mit dem LLTM werden die Aufgabenschwierigkeiten jeder Testaufgabe geschätzt. So kann eindeutig bestimmt werden, für welche Testaufgaben die Schätzung der Aufgabenschwierigkeiten durch die Aufgabenmerkmale genau oder weniger genau erfolgte. Daraus können weiterführende Schlüsse gezogen werden, die in den Abschnitten 8.3.4 und 8.3.5 genutzt und dargestellt werden.

Das LLTM unterscheidet sich vom Rasch-Modell in der Hinsicht, dass die Aufgabenschwierigkeiten durch eine Linearkombination aus mehreren Aufgabenmerkmalen (in der statistischen Terminologie: Basisparametern) bestimmt werden. Aus der Formulierung der Aufgabenschwierigkeit β_j als Linearkombination aus mehreren Basisparametern ergibt sich folgende Erweiterung des Rasch-Modells nach Fischer (1973):

$$\eta_{pi} = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik}$$

Dabei ist X_{ik} die ermittelte Schwierigkeit des Basisparameters für ein Item i des Basisparameters k und β_k indiziert die Gewichtung des Basisparameters k. In Anbetracht des in Abschnitt 8.2.1 formulierten Rasch-Modells wird deutlich, dass

der Itemparameter j durch die lineare Funktion ersetzt wurde:

$$\beta'_j = \sum_{k=0}^{K} \beta_k X_{ik}$$

Dabei entspricht β'_j nicht β_j, da die Schätzung der Aufgabenschwierigkeit durch das LLTM nie der Schätzung des Rasch-Modells perfekt entsprechen wird (Wilson & Boeck, 2004).

Die eingebrachten Basisparameter durch die Erweiterung des Rasch-Modells durch das LLTM werden inhaltlich als die Schwierigkeit von Aufgabenmerkmalen bzw. Aufgabeneigenschaften – im Fall dieser Arbeit der in Kapitel 7 extrahierten Faktoren – der Testaufgaben gedeutet. Die Aufgabenmerkmale werden dahingehend häufig als *kognitive Operatoren* bezeichnet, die benötigt werden, um eine bestimmte Testaufgabe mit der Festlegung einer bestimmten, a priori gesetzten Gewichtung zu lösen (Baghaei & Kubinger, 2015). Damit wird die Aufgabenschwierigkeit über die Anzahl und Art der für die Lösung notwendigen Teiloperationen festgelegt (Zimmermann, 2016).

Der Unterschied zwischen Rasch-Model und LLTM ergibt sich aus dem Beitrag der jeweiligen Aufgabenmerkmale für jede Aufgabe (Wilson & Moore, 2011). Im LLTM wird daher postuliert, dass die Schwierigkeitsparameter vollständig durch die Aufgabenmerkmale erklärt werden können, was das LLTM zu einem restriktiven Modell macht (Wilson & Moore, 2011). Aufgrund der Restriktion wird das LLTM als konservativ erachtet, da häufig nur eine weniger genaue Passung zum Rasch-Modell erreicht wird (Baghaei & Kubinger, 2015; Hartig, 2007; Hartig et al., 2012; Hartig & Frey, 2012; Isaac & Hochweber, 2011; Wilson & Moore, 2011).

8.3.2 Ergebnisse: Modellvergleich

Für das LLTM ergibt sich eine Conditional Log Likelihood von –35230.45 mit einer Parameteranzahl (Anzahl der Faktoren) von $N = 5$ ($SD = 0.65$, $\beta_{\text{MinLLTM}} = -1.7$, $\beta_{\text{MaxLLTM}} = 0.3$). Das Modell ist aufgrund der hohen Parameterreduktion im Vergleich zum Rasch-Modell deutlich sparsamer, wodurch sich jedoch auch die Gesamtvarianz der Aufgabenschwierigkeit reduziert. Insgesamt ergibt sich eine Verringerung der Parameteranzahl um 89.36 %.

Zur Prüfung der Modellgüte des LLTM im Vergleich zum Rasch-Modell wurde die negative doppelte Log-Likelihood verwendet ($\chi^2 = 16930.5$, $df = 41$, p

<.001). Das Ergebnis zeigt eine weniger genaue Passung der durch das LLTM geschätzten Aufgabenschwierigkeit im Vergleich zu den Aufgabenschwierigkeiten, die durch das Rasch Modell geschätzt worden sind – da p signifikant ist, ist ein Unterschied vorhanden der nicht nur zufällig ist. Wie in Abschnitt 8.3.1 erläutert, handelt es sich beim LLTM um ein restriktives Modell; daher war die weniger genaue Passung zu erwarten, insbesondere im Hinblick auf die deutliche Parameterreduktion.

Um zu prüfen, inwieweit die Ergebnisse des LLTM bei der hohen Reduktion der Parameter, die die Aufgabenschwierigkeiten schätzen, als belastbar zu erachten sind, wurden die Beziehungen der geschätzten β-Werte des LLTM mit den Ergebnissen aus dem Rasch-Modell verglichen. In Abbildung 8.5 sind zwei Möglichkeiten des Modellvergleichs zwischen LLTM und Rasch-Modell abgebildet.

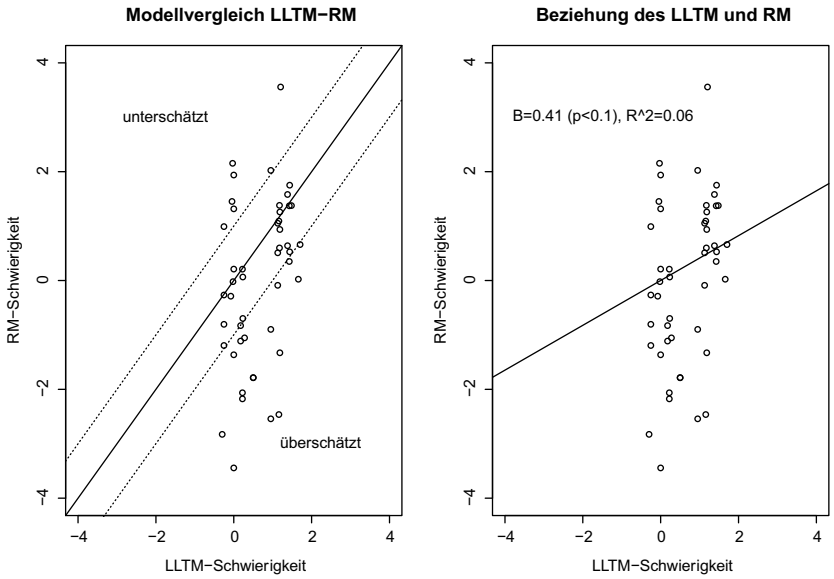

Abbildung 8.5 Modellvergleich des linear-logistischen Testmodells (LLTM) und des Rasch-Modells (RM) – links über eine Winkelhalbierende und rechts durch eine Regression. (Eigene Darstellung)

Die erste Möglichkeit, um die Beziehung zwischen den geschätzten β-Werten zu beurteilen, ist eine grafische Prüfung durch eine Winkelhalbierung des Koordinatensystems zwischen der Traitausprägung des Rasch-Modells und des LLTM. Die Winkelhalbierende indiziert, in welchem Bereich die Aufgabenschwierigkeiten durch das LLTM korrekt geschätzt oder unter- oder überschätzt werden. Diese Möglichkeit ist in Abbildung 8.5 auf der linken Seite dargestellt, mit der Überschrift *Modellvergleich LLTM-RM*. Die zweite Möglichkeit ist der Vergleich der linearen Beziehungen zwischen den geschätzten β-Werten aus beiden Modellen. Diese Möglichkeit ist in Abbildung 8.5 auf der rechten Seite abgebildet, mit der Überschrift *Beziehung des LLTM und RM*.

Auf der linken Seite der Abbildung 8.5 kennzeichnet die Diagonale eine optimale Modellpassung zwischen LLTM und Rasch-Modell. Der Abstand der β-Werte (abgebildet durch die Punkte) zu der Winkelhalbierenden markiert die nicht durch die Basisparameter (Faktoren) erklärte Varianz in den Aufgabenschwierigkeiten. Damit kennzeichnet die Diagonale eine optimale Modellpassung zwischen LLTM und Rasch-Modell. Die gestrichelten, parallel laufenden Linien zu der Diagonale kennzeichnen den Bereich, in dem die Aufgabenschwierigkeit auf ± 0.75 Logits geschätzt wird. Oberhalb der Diagonale werden die Testaufgaben durch das LLTM unterschätzt, unterhalb überschätzt. Beim Modellvergleich zwischen LLTM und RM ist zu erkennen, dass durch das LLTM sieben Testaufgaben (die β-Werte über der gestrichelten Linie) über dem Wert von ± 0.75 Logits unterschätzt werden, während 15 Testaufgaben unter dem Niveau von ± 0.75 Logits überschätzt werden (die β-Werte unter der gestrichelten Linie). Die weiteren 24 Testaufgaben können durch das LLTM auf ± 0.75 Logits genau geschätzt werden (die β-Werte zwischen beiden gestrichelten Linien).

Auf der rechten Seite der Abbildung 8.5 ist die lineare Beziehung zwischen dem LLTM und dem Rasch-Modell dargestellt. Für die durch das LLTM und das Rasch-Modell bestimmte Aufgabenschwierigkeit ergibt sich eine Beziehung von $B = 0.41$ ($t(46) = 1.745$; $p < .1$). Die durch das LLTM geschätzte Aufgabenschwierigkeit zeigt damit nur auf einem Signifikanzniveau von $p < .1$ einen signifikanten Zusammenhang zu den durch das Rasch-Modell geschätzten Aufgabenschwierigkeiten. Für die Beziehung zwischen der Aufgabenschwierigkeit von beiden Modellen ergibt sich ein $R^2 = 0.06$ und damit besitzt die geschätzten Aufgabenschwierigkeiten des LLTM eine geringe Anpassungsgüte (Cohen, 1988). Insgesamt können 6 % der Varianz in den durch das Rasch-Modell bestimmten Aufgabenschwierigkeiten durch die geschätzten Aufgabenschwierigkeiten des LLTM erklärt werden. Daraus ergibt sich ein signifikanter Anteil an Varianz, der durch die Aufgabenschwierigkeiten des LLTM erklärt werden kann ($F(1, 45) = 6.43$; $p < .05$).

Für die durch das LLTM geschätzten Parameter kann eine signifikante Beziehung und Erklärung der Varianz der Aufgabenschwierigkeiten gezeigt werden. Außerdem können die meisten Aufgabenschwierigkeiten durch das LLTM mit einer Genauigkeit von ± 0.75 Logits bestimmt werden. Anhand der dargestellten Ergebnisse kann die Erklärungsleistung des LLTM, im Kontext seiner deutlichen Parameterreduktion und des intendierten Ziels, als ausreichend aussagekräftig betrachtet werden. Aus diesen Gründen kann das LLTM zur weiteren Analyse der Aufgabenschwierigkeiten und des Effektes der Faktoren auf die Aufgabenschwierigkeit genutzt werden. Relevant scheint hierbei jedoch, die Ergebnisse des LLTM im Hinblick auf die Schätzungen der Aufgabenschwierigkeiten je Faktor näher zu betrachten (vgl. Abschnitt 8.3.4).

8.3.3 Ergebnisse: Effekt der Faktoren auf die Aufgabenschwierigkeit

Neben der Ermittlung der Aufgabenschwierigkeiten durch das LLTM kann der Effekt der einzelnen Faktoren auf die Aufgabenschwierigkeit geschätzt werden. Die Faktoren werden als Aufgabenmerkmale behandelt, die zur Lösung der Aufgabe notwendig sind.

Durch die Textmerkmale, die je Beziehung zum Faktor gruppiert wurden, lassen sich unterschiedliche Effekte auf die Textschwierigkeit abschätzen. Durch die in den Abschnitten 2.4.3 und 5.4.2 dargestellten theoretischen Erkenntnisse kann die Annahme getroffen werden, dass sich die Textschwierigkeit positiv auf die Aufgabenschwierigkeit auswirken kann. Die Ergebnisse der η-Schätzungen für den Effekt auf die Schwierigkeit der Aufgaben sind in Abbildung 8.6 dargestellt.

In der Abbildung 8.6 ist jeweils zu erkennen, dass der erklärende Faktor den stärksten positiven Effekt mit $\eta = 1.204$ auf die Aufgabenschwierigkeit aufweist. Weitere positive, jedoch deutlich geringere Effekte auf die Aufgabenschwierigkeiten werden für den informativen ($\eta = 0.222$) und den instruktiven ($\eta = 0.277$) Faktor durch das LLTM geschätzt. Einen negativen Effekt auf die Aufgabenschwierigkeit haben der komprimierende ($\eta = 0.252$) und beschreibende ($\eta = 0.045$) Faktor, was bedeutet, dass die Aufgaben durch die Faktoren leichter werden.

Um die Ergebnisse im Hinblick auf die gemeinsam vorkommenden Textmerkmale und deren Effekt auf die Textschwierigkeit zu deuten, erfolgt an dieser Stelle eine Einordnung, die als Diskussionsgrundlage für das Abschnitt 8.4 dient.

Erklärend: Der deutliche positive Effekt auf die Aufgabenschwierigkeit zeigt sich bei den erklärenden Faktor. Der erklärende Faktor vereinigt Textmerkmale,

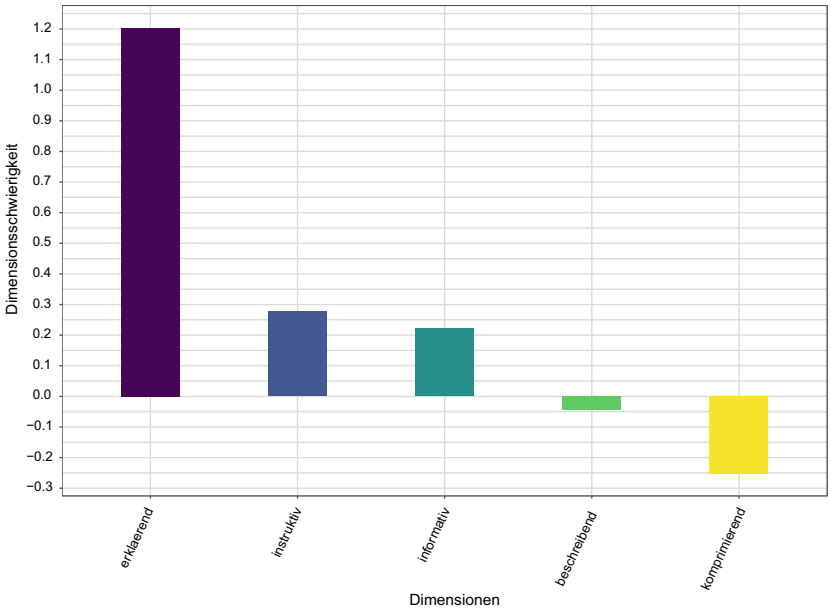

Abbildung 8.6 Effekte der Faktoren auf die Aufgabenschwierigkeiten (η). (Eigene Erstellung)

die sich positiv auf die Textschwierigkeit auswirken können. Das betrifft insbesondere die Verwendung von mathematischen Begriffen als Fachtermini für den Mathematikunterricht sowie die Nutzung von Nominalisierung und unpersönlicher Sprache (vgl. Abschnitt 5.2.3). Jedoch haben auch Textmerkmale eine hohe Bedeutung für den erklärenden Faktor, die auch einen negativen Effekt auf die Textschwierigkeit aufweisen können. Dazu zählen Füllwörter zur Schaffung zusätzlicher Redundanz, die – wie in Abschnitt 5.3.3 erläutert – insbesondere für fachliche Texte bedeutsam sein können, die Verwendung von Konjunktionen und Präpositionen zur Entwicklung von Textkohäsion und die Verwendung von Zahlen als schnell ablesbare Informationen (vgl. Abschnitt 5.2.3 und Abschnitt 5.2.4). Die Ergebnisse sprechen dafür, dass Kohäsionsmittel in Verbindung mit einer häufigen Verwendung von begrifflichen Mitteln einen Effekt auf die Aufgabenschwierigkeit haben.

Instruktiv: Den zweitstärksten positiven Effekt auf die Aufgabenschwierigkeit hat der instruktive Faktor. Dieser zeichnet sich durch Symbole und mathematische

Begriffe als charakteristische Textmerkmale aus. Sowohl die Verwendung von Symbolen als auch die Nutzung von mathematischen Begriffen können bereits separat betrachtet als sich positiv auf die Textschwierigkeit auswirkend beurteilt werden (vgl. Abschnitt 5.2.3 und Abschnitt 5.2.4). Die Ergebnisse des Effektes auf die Aufgabenschwierigkeit unterstreichen diesen separaten Effekt bei dem systematischen gemeinsamen Vorkommen dieser Textmerkmale.

Informativ: Ebenfalls zeigt der informative Faktor einen positiven Effekt auf die Aufgabenschwierigkeit. Dieser fällt etwas geringer aus als der bereits moderate Effekt für den instruktiven Faktor. In Anbetracht der Textmerkmale, die in diesem Faktor gemeinsam vorkommen, lassen sich unterschiedliche Effekte auf die Textschwierigkeit abschätzen. So soll sich die Gebräuchlichkeit des Wortschatzes negativ und ein hoher propositionaler Gehalt tendenziell positiv auf die Textschwierigkeit auswirken. Der informative Faktor zeigt insgesamt einen negativen Effekt auf die Aufgabenschwierigkeit. Die Ergebnisse geben einen Hinweis darauf, dass der hohe propositionale Gehalt einen positiven Effekt auf die Aufgabenschwierigkeit hat. Die Verwendung von gebräuchlichem Wortschatz nimmt in diesem Zusammenhang keinen großen Einfluss auf die Aufgabenschwierigkeit.

Beschreibend: Der beschreibende Faktor kann einen schwachen negativen Effekt auf die Aufgabenschwierigkeit aufweisen. Für den beschreibenden Faktor sind besonders Text-Text-Referenzen charakteristisch. Direkte Anaphorik sollte tendenziell zu einer leichten Kohärenzbildung führen und damit zu einer geringeren Textschwierigkeit (vgl. Abschnitt 5.2.3 und 5.2.4). Die lexikalische Vielfalt indiziert den Gebrauch von unterschiedlichem Vokabular; damit kann mindestens eine Erhöhung des Wortschatzes einhergehen.

Nicht charakteristisch sind für den beschreibenden Faktor mathematische Begriffe, die wie für den erklärenden Faktor erläutert, einen positiven Effekt auf die Textschwierigkeit aufweisen können. Diskontinuierlicher Text sollte sich generell erleichternd auf die Textschwierigkeit auswirken (vgl. Abschnitt 5.3.4). Das Textmerkmal diskontinuierlicher Text ist jedoch nicht charakteristisch für diesen Faktor.

Dahingehend deuten die Ergebnisse darauf hin, dass die Verwendung von Textkohärenzstrukturen und die Reduktion der Verwendung von mathematischen Begriffen insgesamt zu einem leicht negativen Effekt auf die Aufgabenschwierigkeit führen.

Komprimiert: Der komprimierende Faktor kann einen deutlichen negativen Effekt auf die Aufgabenschwierigkeit aufweisen – und dass, obwohl in diesem Faktor viele Textmerkmale (Passiv, durchschnittliche Silbenanzahl) vorhanden sind, die einen positiven Effekt auf die Textschwierigkeit haben sollten (Abschnitt 5.2.3 und Abschnitt 5.2.4). Ein Textmerkmal, das sich tendenziell

negativ auf die Textschwierigkeit auswirken soll, ist der diskontinuierliche Text (vgl. Abschnitt 5.3.4). Die Ergebnisse weisen darauf hin, dass unterschiedliche Formen von Darstellungen zu einem negativen Effekt auf die Aufgabenschwierigkeit führen können – trotz der auch gemeinsam verwendeten Textmerkmale wie einer durchschnittlichen Silbenanzahl, Passiv und lexikalischer Vielfalt.

Mit der Ermittlung des Effektes der Faktoren auf die Aufgabenschwierigkeit wurde die vierte Zielvoraussetzung für die Erstellung des Instruments zur sprachlichen Variation von Textaufgaben geleistet. Durch die empirisch ermittelten Effekte können bedeutende Implikationen für Anpassungsmöglichkeiten von Textaufgaben an Lernende getroffen werden, die in Abschnitt 8.4 diskutiert werden.

8.3.4 Ergebnisse: Vergleich der geschätzten Aufgabenschwierigkeit des linear-logistischen Testmodells und des Rasch-Modells

In Abschnitt 8.3.2 wurde verdeutlicht, dass das LLTM bestimmte Aufgaben in Vergleich zum Rasch-Modell genau einschätzt, andere Aufgabenschwierigkeiten weniger genau. Im Hinblick auf den in Abschnitt 8.3.3 dargestellten Effekt der Faktoren auf die Aufgabenschwierigkeiten stellt sich die Frage, inwieweit gewisse Faktoren besonders häufig bei einer genauen bzw. weniger genauen Passung ausgeprägt sind und welche Interaktionen gegebenenfalls häufiger sind.

In Abbildung 8.7 ist die absolute Differenz, also der Betrag der Differenz, zwischen den geschätzten Aufgabenschwierigkeiten des LLTM und des Rasch-Modells als Balkendiagramm abgebildet. Die Aufgaben wurden von hoher (lila) nach geringer (gelb) Passung sortiert. Ungefähr die Hälfte der Aufgaben wird mit ± 1 Logits genau geschätzt, die andere Hälfte wird über ± 1 Logits falsch eingeschätzt.

Um zu prüfen, inwieweit bestimmte Faktoren besonders dafür geeignet sind, die Aufgabenschwierigkeiten durch das LLTM einzuschätzen, wurden die geschätzten Aufgabenschwierigkeiten der Testaufgaben zwischen ± 1 Logits getrennt und die Ausprägungen der Faktoren wurde betrachtet. Dabei wurde verglichen, welche Faktoren eine (1) und keine (0) Ausprägung bei den jeweiligen Testaufgaben hatten. Dadurch konnte ermittelt werden, ob das Auftreten bestimmter Faktoren zu einer ungenaueren oder genaueren Einschätzung der Aufgabenschwierigkeiten führt.

Genaue Schätzung der Aufgabenschwierigkeit: Der prozentuale Anteil der Faktoren mit einer Ausprägung (kodiert durch eine 1, nach Dichotomisierung) liegt

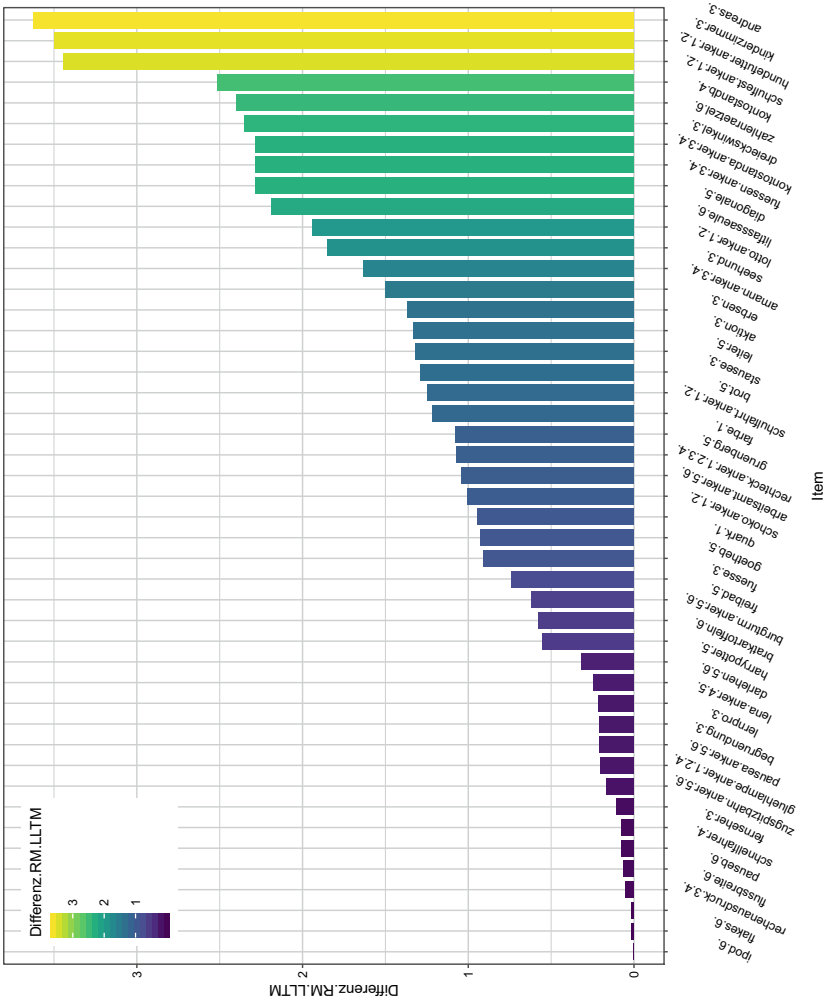

Abbildung 8.7 Differenzbildung zur Verdeutlichung der Passung. (Eigene Erstellung)

bei 48.70 %, für keine Ausprägung (kodiert durch eine 0, nach Dichotomisierung) bei 51.30 %. Für die geschätzten Aufgabenschwierigkeiten nach dem LLTM, die eine hohe Passung zu den Aufgabenschwierigkeiten des Rasch-Modells erreichen, hat der erklärende Faktor einen Anteil von 56.52 % (11.30 %), der komprimierende Faktor einen Anteil von 47.83 % (9.57 %), der beschreibende Faktor einen Anteil von 60.87 % (12.17 %), der informative Faktor einen Anteil von 43.48 % (8.70 %) und der instruktive Faktor einen Anteil von 34.78 % (6.96 %) der Ausprägungen der jeweiligen Faktoren (bzw. aller Ausprägungen sowohl mit 1 als auch 0 als Ausprägung). Den höchsten prozentualen Anteil einer Ausprägung erreicht der beschreibende Faktor mit 25.00 %. Der instruktive Faktor hat mit einem prozentualen Anteil der Ausprägungen von 14.29 % den geringsten Anteil an Ausprägungen auf den Testaufgaben.

Im Durchschnitt ergeben sich 2.435 Ausprägungen für die Faktoren. Die höchsten Ausprägungen (max. = 4) haben die Aufgaben *Darlehen*, *Fernseher*, *Schnellfahrer* und *Freibad*.

Ungenaue Schätzung der Aufgabenschwierigkeit: Für die Aufgabenschwierigkeiten, die durch das LLTM ungenau eingeschätzt werden, erhielten 42.39 % eine Ausprägung und 57.61 % keine Ausprägung. Davon hat der erklärende Faktor einen Anteil von 43.48 % (8.70 %), der komprimierende Faktor einen Anteil von 30.43 % (6.09 %), der beschreibende Faktor einen Anteil von 26.09 % (5.22 %), der informative Faktor einen Anteil von 43.48 % (8.70 %) und der instruktive Faktor einen Anteil von 26.09 % (5.22 %) der Ausprägung der jeweiligen Faktoren (bzw. aller Ausprägungen sowohl mit 1 als auch 0 als Ausprägung). Den höchsten prozentualen Anteil einer Ausprägung erreichen sowohl der erklärende als auch der informative Faktor mit 25.64 %. Der instruktive und der beschreibende Faktor haben mit einem prozentualen Anteil der Ausprägungen von 15.38 % den geringsten Anteil der Ausprägungen. Im Durchschnitt ergeben sich 1.696 Ausprägungen für die Faktoren. Die höchsten Ausprägungen (max. = 4) haben die Aufgaben *Schulfest*, *Schulfahrt* und *Seehund*.

Die Ergebnisse deuten darauf hin, dass die geschätzte Aufgabenschwierigkeit durch das LLTM bei häufigem Auftreten des erklärenden und beschreibenden Faktors sowie geringem Vorkommen des instruktiven Faktors die vom Rasch-Modell ermittelte Aufgabenschwierigkeit genau vorhersagen kann. Im Gegensatz dazu weisen die Ergebnisse darauf hin, dass bei einem hohen Anteil des erklärenden und informativen Faktors und einem geringen Anteil des instruktiven und beschreibenden Faktors die Aufgabenschwierigkeiten durch das LLTM tendenziell ungenauer geschätzt werden.

8.3.5 Ergebnisse: Erklärungsleistung der Faktoren für unterschiedlich schwierige Aufgaben

Neben der in Abschnitt 8.3.4 durchgeführten Betrachtung einer genauen oder weniger genauen Schätzung anhand einer Abweichung von ± 1 Logits ist ebenfalls von Interesse, welche Testaufgaben, die in Abbildung 8.7 dargestellt sind, genau oder weniger genau durch das LLTM geschätzt worden sind. Dahingehend können exemplarisch die in Abbildung 8.7 dargestellten drei am genauesten bzw. am ungenauesten geschätzten Testaufgaben analysiert werden. Unter den Testaufgaben, deren Aufgabenschwierigkeit durch das LLTM genau geschätzt wurde, sind die Testaufgaben *IPod*, *Flakes* und *Rechenausdruck*. Die Testaufgaben, deren Aufgabenschwierigkeit (deutlich) ungenauer geschätzt wurde, sind die Aufgaben *Hundefutter*, *Kinderzimmer* und *Andreas*. Die Testaufgaben sind in Tabelle 8.4 dargestellt.

Bei einer Betrachtung der durch das Rasch-Modell berechneten Aufgabenschwierigkeiten der in Tabelle 8.4 abgebildeten Testaufgaben in den Tabellen 8.1 und 8.2 ist erkennbar, dass die Testaufgaben, die eine ungenaue Schätzung der Aufgabenschwierigkeiten aufweisen, solche Aufgaben sind, für die eine geringe Aufgabenschwierigkeit berechnet wurde. Die Testaufgaben, bei denen eine genaue Schätzung der Aufgabenschwierigkeit durch das LLTM gelungen ist, sind solche, deren Aufgabenschwierigkeit durch das Rasch-Modell höher eingeschätzt wurde. Dies gibt einen Hinweis darauf, dass sich die Passung der Schätzungen der Aufgabenschwierigkeiten des LLTM unterscheidet, je nachdem, welche Aufgabenschwierigkeit durch das Rasch-Modell ermittelt wurde.

In Abbildung 8.8 ist der Vergleich der geschätzten Aufgabenschwierigkeit zwischen dem Rasch-Modell und dem LLTM als Balkendiagramm dargestellt. Die berechneten β-Werte wurden nach dem Rasch-Modell aufsteigend sortiert. Die Abbildung verdeutlicht, dass das LLTM eine hohe Passung der Aufgabenschwierigkeiten erreicht, wenn für die Aufgabenschwierigkeit ein positives β berechnet wird. Bei negativen β der Aufgabenschwierigkeit weichen die Schätzungen der Aufgabenschwierigkeiten durch das LLTM deutlich von den berechneten Aufgabenschwierigkeiten des Rasch-Modells ab.

Um die Erklärungsleistung des LLTM für die unterschiedlichen geschätzten Aufgabenschwierigkeiten zu bestimmen, wurden die Testaufgaben nach den bestimmten Aufgabenschwierigkeiten des LLTM und des Rasch-Modells geclustert. Dieses Vorgehen wurde gewählt, um die nahe Beziehung der Aufgabenschwierigkeiten auf der Skala des LLTM und des Rasch-Modells einzubeziehen, was bei einer diskreten Einteilung durch die z. B. vom Rasch-Modell bestimmten Aufgabenschwierigkeiten nicht möglich wäre.

Tabelle 8.4 Sechs exemplarische Testaufgaben – jeweils drei mit genauer und ungenauer Schätzung der Aufgabenschwierigkeit durch das linear-logistische Testmodell (LLTM)

Genaue Schätzung der Aufgabenschwierigkeit durch das LLTM	Ungenaue Schätzung der Aufgabenschwierigkeit durch das LLTM
Aufgabe IPod: Ein IPod kostet 200 €; dazu kommen 19 % Mehrwertsteuer. Bei Barzahlung reduziert sich dieser Betrag (einschließlich Mehrwertsteuer) um 3 %. Wie viel € muss der Kunde zahlen? Schreibe auf, wie du gerechnet hast.	Aufgabe Andreas: Andreas lädt seine Freunde ein. Er muss 4 Cola zu je 1,20 €, 3 Eisbecher zu je 3,60 € und einen Milchshake bezahlen. Andreas gibt dem Kellner einen 20-€-Schein und erhält 2,30 € zurück. Wie teuer ist der Milchshake? Schreibe auf, wie du gerechnet hast.
Aufgabe Flakes: Auf abgepackten Lebensmitteln muss der Händler neben dem Verkaufspreis auch den Preis für 1 kg angeben. Der folgende Ausschnitt zeigt ein Angebot eines Supermarkts. Überprüfe, ob der 1 kg-Preis stimmt. Je 425-g-Pckg. 1.99 (kg-Preis 4,68)	Aufgabe Kinderzimmer: Anna und Stefan wollen ihre Kinderzimmer neu streichen. Stefan mischt 2 l weiße Farbe mit 5 l gelber Farbe, Anna mischt 1 l weiße Farbe mit 2 l gelber Farbe. Wer erhält die hellere Mischung? Begründe deine Antwort.
Aufgabe Rechenausdruck: Schreibe als Rechenausdruck und berechne ihn: Multipliziere die Summe aus –3 und + 10 mit der Differenz aus –4 und –7.	Aufgabe Hundefutter: Kathrin hat fünf Dosen Hundefutter gekauft. Zusammen kosten die Dosen 10,50 €. Wie viel kostet eine Dose?

Die Ergebnisse der Clustergruppierung der Testaufgaben ist in Abbildung 8.9 dargestellt. Für das Gruppieren der Variablen wurden drei Cluster vorgewählt und ein hierarchisches Clustern genutzt. In der Abbildung ist erkennbar, dass sich drei Gruppen besonders auf der Skala der Schwierigkeit des Rasch-Modells abbilden lassen.

Das erste Cluster, in Abbildung 8.9 lila dargestellt, sind Testaufgaben, die eine Aufgabenschwierigkeit nach dem Rasch-Modell im unteren Bereich der Skala aufweisen. Die Testaufgaben können als die leichtesten Aufgaben gedeutet werden. Im zweiten Cluster, in Abbildung 8.9 blaugrün abgebildet, werden die Testaufgaben zusammengefasst, die nach dem Rasch-Modell im mittleren Bereich der Skala der Ausprägungen der Aufgabenschwierigkeiten liegen. Diese Testaufgaben können als mittelschwierig interpretiert werden. Das dritte Cluster, in Abbildung 8.9 in Gelb dargestellt, sind Testaufgaben, die eine Aufgabenschwierigkeit im oberen Bereich aufweisen. Die Testaufgaben in diesem Bereich

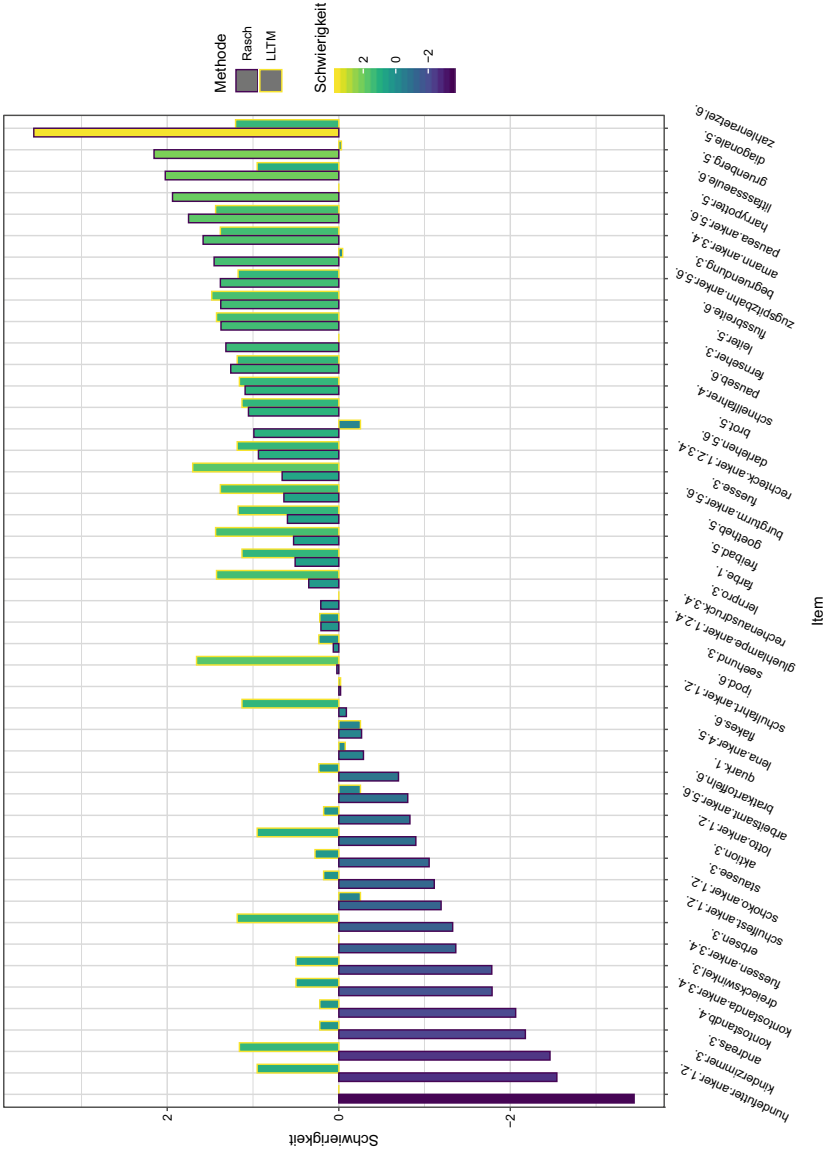

Abbildung 8.8 Vergleich der geschätzten Aufgabenschwierigkeit je Testaufgabe. (Eigene Erstellung)

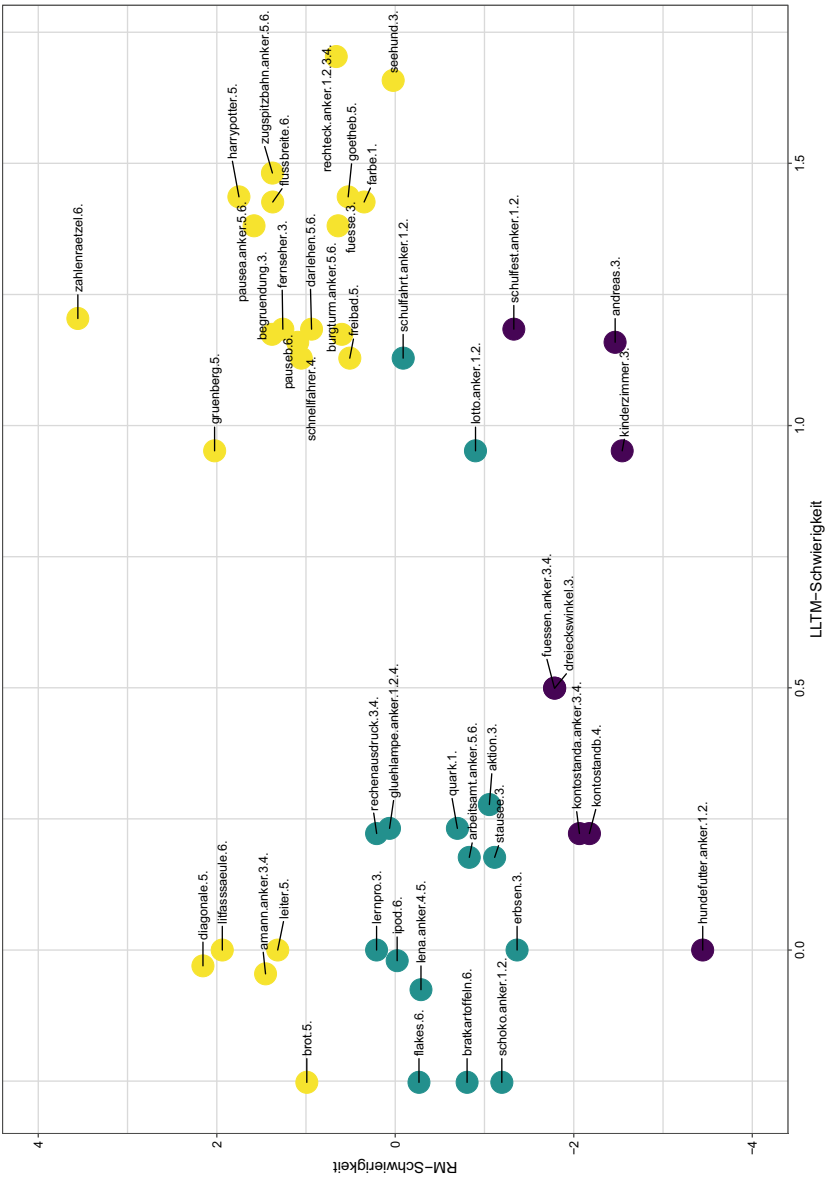

Abbildung 8.9 Gruppenbildung von Textaufgaben durch hierarchisches Clustern nach Schwierigkeit der Rasch-Analyse und geschätzter Schwierigkeit nach linear-logistischem Testmodell (LLTM). (Eigene Erstellung)

können als schwierig betrachtet werden. Darüber hinaus ist in Abbildung 8.9 zu erkennen, dass je Cluster auf der Skala der Aufgabenschwierigkeit nach dem LLTM kaum eine Differenzierung existiert. Die deutlich höhere Varianz der Aufgabenschwierigkeiten des Rasch-Modells bestimmt die Gruppenbildung.

Die drei Cluster an Testaufgaben werden als Testaufgaben mit unterschiedlichen Aufgabenanforderungen betrachtet. Um nun die Erklärungsleistung der Faktoren (Aufgabenmerkmale) im Hinblick auf die Aufgabenanforderung zu ermitteln, wurden wie in Abschnitt 8.3.2 ein Modellvergleich zwischen LLTM und Rasch-Modell durch eine Winkelhalbierung des Koordinatensystems zwischen den abgetragenen Aufgabenschwierigkeiten beider Modelle sowie eine Regressionsanalyse durchgeführt.

In Abbildung 8.10 sind die Ergebnisse des Modellvergleichs durch Winkelhalbierung dargestellt. Die zur Winkelhalbierenden parallel laufenden beiden Geraden markieren eine Abweichung von ± 0.75 Logits (vgl. Abschnitt 8.3.2). Auf der linken Seite der Abbildung 8.10 sind die Gruppen der leichten Aufgaben abgetragen. In der Abbildung ist zu erkennen, dass alle Aufgabenschwierigkeiten für leichte Aufgaben durch das LLTM überschätzt werden. Das bedeutet, dass die durch die Faktoren (als Aufgabenmerkmale) ermittelte Aufgabenschwierigkeit höher ist als die von dem Rasch-Modell ermittelte.

In der Mitte der Abbildung 8.10 sind die Schätzungen der Aufgabenschwierigkeit der mittelschwierigen Aufgaben gegeneinander aufgezeichnet. Für diese Aufgaben ist die Passung zwischen den geschätzten Aufgabenschwierigkeiten von beiden Modellen genauer als für leichte Aufgaben. In Abbildung 8.10 ist zu erkennen, dass die 13 mittelschwierigen Testaufgaben durch das LLTM im Bereich von ± 0.75 Logits geschätzt werden können. Lediglich fünf Testaufgaben werden durch das LLTM für mittelschwierige Testaufgaben überschätzt.

Auf der rechten Seite der Abbildung 8.10 sind die Schätzungen der Aufgabenschwierigkeiten der schweren Aufgaben gegeneinander aufgetragen. Auch für schwierige Aufgaben zeigt sich ein genaues Ergebnis zur Schätzung der Aufgabenschwierigkeiten durch das LLTM. Die Schätzungen von 14 Testaufgaben erfolgen im Rahmen von ± 0.75 Logits. Nur noch zwei Testaufgaben werden überschätzt und acht Aufgaben werden unterschätzt.

Die Ergebnisse der Unterscheidung des LLTM für unterschiedliche Aufgabenanforderungen deuten darauf hin, dass sprachliche Faktoren gut zur Schätzung der Aufgabenschwierigkeit dienen können, wenn die Testaufgaben eine mittlere oder hohe Aufgabenanforderung aufweisen. Für leichte Testaufgaben wird die Aufgabenschwierigkeit durch sprachliche Faktoren generell überschätzt.

Der Modellvergleich durch die Winkelhalbierung lässt keine Rückschlüsse bezüglich eines quantitativen Indikators eines Erklärungseffektes zu. Aus diesem

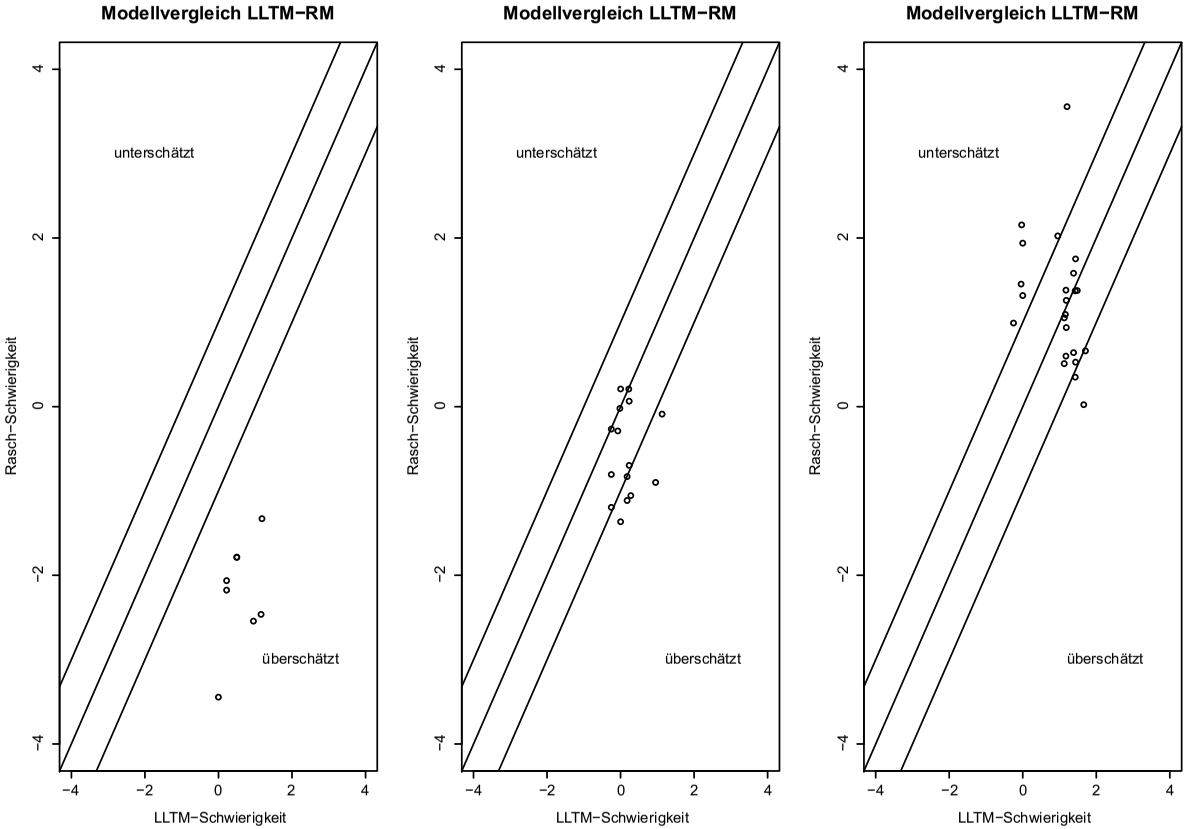

Abbildung 8.10 Modellvergleich der Passung der geschätzten Aufgabenschwierigkeiten durch die Winkelhalbierende nach Aufgabenanforderung (links: leichte Aufgaben, Mitte: mittelschwierige Aufgaben, rechts: schwierige Aufgaben). (Eigene Erstellung)

Grund wurde wie in Abschnitt 8.3.2 für den Modellvergleich eine Regressionsanalyse durchgeführt, um so die Erklärungsleistung des LLTM für die unterschiedlichen Aufgabenschwierigkeiten zu quantifizieren.

Für die Regressionsanalyse wurden die Aufgabenschwierigkeiten nach den Ergebnissen der Clusterung getrennt analysiert. In Abbildung 8.11 und in Tabelle 8.5 sind die Ergebnisse der Regressionsanalyse für die Aufgabenschwierigkeiten beider Modelle abgebildet.

In Tabelle 8.5 ist zu erkennen, dass die Schätzungen der Aufgabenschwierigkeiten durch das LLTM für leichte Aufgaben, die in Abbildung 8.11 auf der linken Seite abgebildet sind, mit einem negativen Effekt ($B = 0.116$; $t(7) = 2.751$; $p < .05$) signifikant werden. Die durch das LLTM geschätzte Aufgabenschwierigkeit für leichte Aufgaben kann einen hohen Anteil an Varianz der Aufgabenschwierigkeit der durch das Rasch-Modell ermittelten Aufgabenschwierigkeit erklären (52 %).

Für mittelschwierige Testaufgaben (in Abbildung 8.11 in der Mitte dargestellt), kann das LLTM die durch das Rasch-Modell bestimmte Aufgabenschwierigkeit nicht signifikant voraussagen ($B = 0.41$; $t(14) = 0.885$; (ns)). Der Anteil der erklärten Varianz der durch das LLTM geschätzten Aufgabenschwierigkeiten kann lediglich 5.29 % der Varianz der Aufgabenschwierigkeit des Rasch-Modells erklären. Im Vergleich zum Modellvergleich durch die Winkelhalbierung sind die Ergebnisse der Regressionsanalyse deutlich restriktiver und die Güte der Passung ist negativer zu beurteilen.

Die Schätzungen der Aufgabenschwierigkeiten des LLTM bei schwierigen Aufgaben ist in Abbildung 8.11 rechts dargestellt und in Tabelle 8.5 zusammengefasst. Hinsichtlich der schwierigen Testaufgaben kann das LLTM die Aufgabenschwierigkeit signifikant vorhersagen ($R = 0.81$; $t(22) = 4.14$; $p > .001$). Ein hoher Anteil der Varianz der Aufgabenschwierigkeit bei schwierigen Aufgaben kann durch die geschätzten Aufgabenschwierigkeiten des LLTM erklärt werden (43.79 %).

Durch die Regressionsanalyse können die Ergebnisse des ersten Modellvergleichs durch die Winkelhalbierung über quantitative Indikatoren verändert eingeschätzt werden. So können zwar für mittelschwierige Aufgaben die meisten Testaufgaben in einem Bereich von ± 0.75 Logits genau geschätzt werden, doch die Ergebnisse der Regressionsanalyse konnten keine robuste Erklärungsleistung der geschätzten Aufgabenschwierigkeiten des LLTM darstellen. Für leichte Aufgaben zeigen die Ergebnisse einen negativen Effekt auf die geschätzten Aufgabenschwierigkeiten. Dies deutet darauf hin, dass das Vorkommen von Faktoren einen Einfluss auf die Aufgabenschwierigkeit zeigt, die Aufgabenleichtigkeit kann

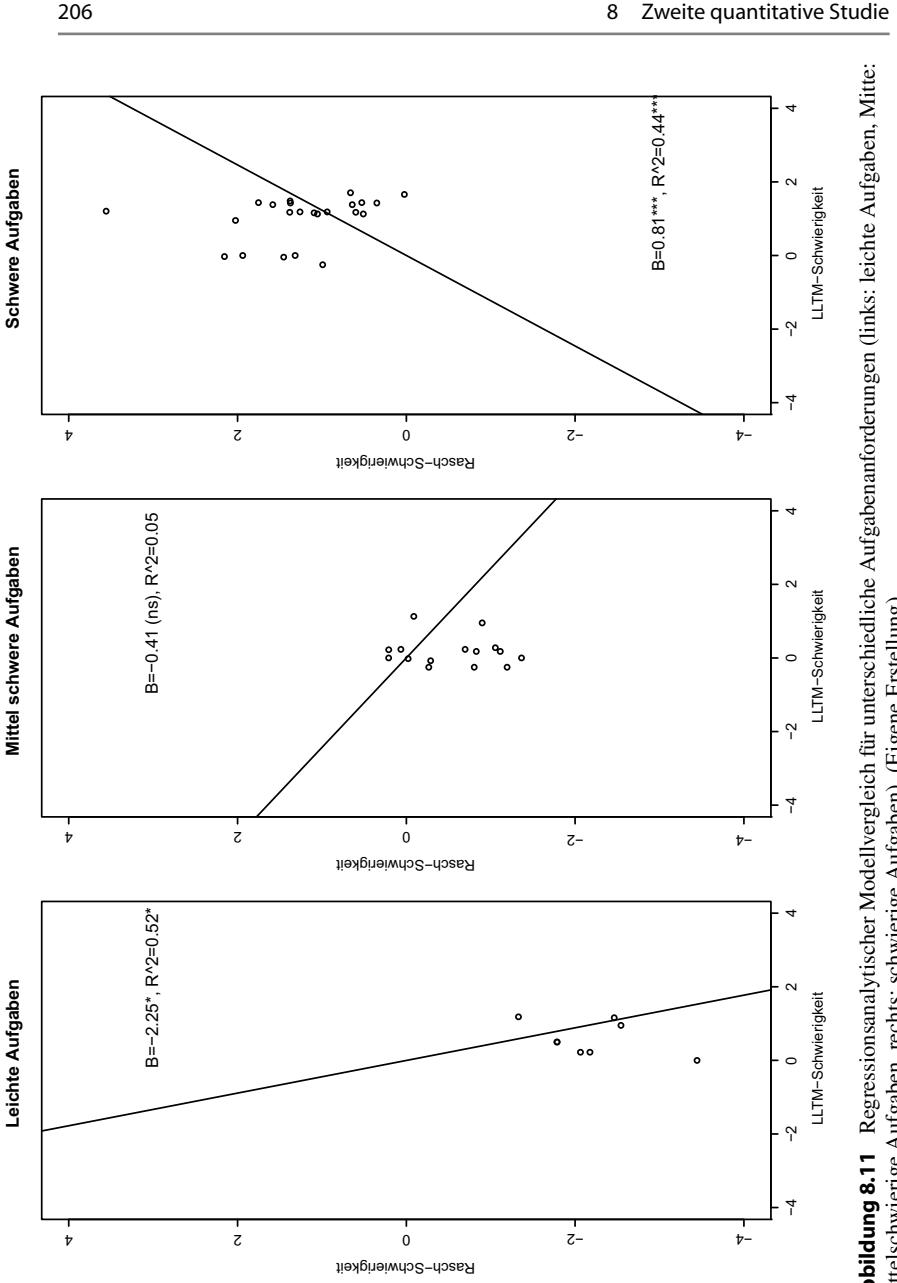

Abbildung 8.11 Regressionsanalytischer Modellvergleich für unterschiedliche Aufgabenanforderungen (links: leichte Aufgaben, Mitte: mittelschwierige Aufgaben, rechts: schwierige Aufgaben). (Eigene Erstellung)

Tabelle 8.5 Ergebnisse der Regressionsanalyse der geschätzten Aufgabenschwierigkeiten des linear-logistischen Testmodells (LLTM) und des Rasch-Modells (RM) für unterschiedliche Aufgabenanforderungen

Leichte Aufgaben	Prädikator: LLTM-Aufgabenschwierigkeiten	
B	– 0.253*	KI 95 % [–4.190, –0.316]
R^2	0.519*	
AIC	34.02	
BIC	34.17	
Mittelschwierige Aufgaben		
B	– 0.411	KI 95 % [–1.407, 0.585]
R^2	0.027	
AIC	37.27	
BIC	38.69	
Schwierige Aufgaben		
B	0.813***	KI 95 % [0.406, 1.220]
R^2	0.440***	
AIC	72.88	
BIC	75.15	

*$p < 0.5$; **$p < 0.01$; ***$p < 0.00$

jedoch nicht ausreichend durch die Faktoren modelliert werden. Die hohe Erklärungsleistung der geschätzten Aufgabenschwierigkeiten für schwierige Aufgaben weist ebenfalls darauf hin, dass die Schätzung der Aufgabenschwierigkeiten durch die Faktoren als Aufgabenmerkmale gelingt. Mit den Faktoren des Instruments zur sprachlichen Variation von Textaufgaben im Mathematikunterricht kann damit die Aufgabenschwierigkeit für schwierige Textaufgaben genau geschätzt werden.

8.4 Diskussion

Die Ergebnisse des Rasch-Modells und des LLTM bieten eine Reihe von Ergebnissen, die auf unterschiedlichen Ebenen für das Ziel der Konzeptualisierung eines Instruments zur sprachlichen Veränderung diskutiert werden können. Die Diskussion wird aus diesem Grund in drei generelle Ebenen unterteilt:

1. Modellvergleich von LLTM und Rasch-Modell
2. Effekte der Faktoren auf die Aufgabenschwierigkeit
3. Modellvergleich von LLTM und Rasch-Modell im Hinblick auf unterschiedliche Aufgabenanforderungen

Modellvergleich: Zur Schätzung der durch das Rasch-Modell bestimmten Aufgabenschwierigkeit wurden im LLTM die sprachlichen Faktoren als Aufgabenmerkmale verwendet. Durch das LLTM wurden die geschätzten Aufgabenschwierigkeiten des Rasch-Modells zu einem bedeutenden Teil korrekt eingeschätzt. Des Weiteren hat das LLTM im Regressionsmodell einen noch signifikanten Effekt auf die Erklärungsleistung der Aufgabenschwierigkeiten des Rasch-Modells – jedoch nur bei einer geringen Varianzaufklärung.

Die Ergebnisse des Modellvergleichs bedeuten für das Instrument zur sprachlichen Variation, dass die sprachlichen Faktoren des Instruments insoweit dazu geeignet sind, die Aufgabenschwierigkeiten vorherzusagen, als dass die schwierigkeitsgenerierende Effekte der sprachlichen Faktoren betrachtet werden können. Die Befunde deuten darauf hin, dass durch die Veränderung der Faktoren die Anpassung zwischen Text und Lesenden gelingen kann, da es möglich ist, durch die Faktoren die Aufgabenschwierigkeit zu schätzen(vgl. Abschnitt 2.4.3 und 6.1). Die geringe Varianzaufklärung zeigt jedoch, dass weitere Faktoren (z. B. inhaltliche) für die Aufgabenlösung eine ebenfalls hohe Bedeutung aufweisen. Die eingeschränkte Erklärungsleistung des LLTM war zu erwarten, da bereits in Kapitel 5 die Eingrenzung der Untersuchung nur auf Textfaktoren erörtert wurde, bei der die Interaktion von Text und Rezipientinnen bzw. Rezipienten nicht mitbetrachtet wurde. So werden das Vorwissen der Rezipientinnen und Rezipienten sowie deren allgemeine sprachliche Fähigkeiten nicht durch die Textmerkmale berücksichtigt. Außerdem werden durch die sprachlichen Faktoren des Instruments nicht die inhaltlich-konzeptuellen Fähigkeiten erhoben, die notwendig sind, um die Aufgaben zu lösen. Aus diesem Grund wurde zur Spezifizierung der Erklärungsleistung der geschätzten Aufgabenschwierigkeit der sprachlichen Faktoren nach unterschiedlichen Anforderungsgruppen unterschieden (vgl. Punkt 3 der Diskussionsebenen).

Effekte der Faktoren auf die Aufgabenschwierigkeit: Die geschätzten Effekte der Faktoren zeigen sowohl einen positiven als auch negativen Einfluss auf die Aufgabenschwierigkeit. Der erklärende, der instruktive und der informative Faktor zeigen positive Effekte auf die Aufgabenschwierigkeit, während der beschreibende und der komprimierende Faktor einen negativen Effekt auf die Aufgabenschwierigkeit aufweisen. Im Folgenden sollen die Ergebnisse für die unterschiedlichen Faktoren gedeutet werden:

Erklärend: Der hohe negative Effekt auf die Aufgabenschwierigkeit des erklärenden Faktors ist durch die begrifflich dominanten Texte aufgrund der häufigen Verwendung von Fachtermini und Nominalisierung sowie unpersönlicher Sprache zu erklären. Ein Grund hierfür kann die mit der Verwendung von Begriffen einhergehende Bedeutung des konzeptuellen Verständnisses dieser Begriffe sein, so ist diesbezüglich die kognitive Funktion von Sprache und der Aspekt der Sprache als Lerngegenstand für den Mathematikunterricht zu nennen (vgl. Abschnitt 2.3 und Abschnitt 2.4.1). Ergänzend kann die logische Verknüpfung von Begriffen durch Konjunktionen und Präpositionen für mathematische Textaufgaben einen herausfordernden Charakter aufweisen (vgl. Abschnitt 5.2.4). Die Verknüpfung des Einflusses der typischen Textmerkmale des erklärenden Faktors kann damit aufgrund der Bedeutung von Begriffen und der Verwendung von Funktionswörtern für den mathematischen Gegenstand unter der Perspektive einer kognitiven Funktion von Sprache auf der Ebene der Textschwierigkeit sowohl indirekt einen Einfluss auf die Aufgabenschwierigkeit haben als auch direkt.

Instruktiv: Für den instruktiven Faktor ergibt sich ein positiver Effekt auf die Aufgabenschwierigkeiten. Aufgrund der Textmerkmale, die für diesen Faktor spezifisch sind, entspricht das Ergebnis den Erwartungen, die durch den Einfluss auf die Textschwierigkeit bestehen. Der positive Effekt auf die Aufgabenschwierigkeiten lässt sich neben dem indirekten Einfluss durch die Erhöhung der Textschwierigkeit, wie für den erklärenden Faktor, mit der Verbindung der kognitiven Funktion von Sprache und dem Aspekt der Sprache als Lerngegenstand interpretieren (vgl. Abschnitt 2.3 und Abschnitt 2.4.1). Mathematische Begriffe, aber auch (mathematische) Symbole, erfordern ein konzeptuelles Verständnis. Aufgrund dessen ist die kognitive Funktion dieser Textmerkmale besonders bedeutend und verweisen ebenfalls auf Sprache als Lerngegenstand im Mathematikunterricht.

Informativ: Der informative Faktor hat insbesondere zwei charakteristische Textmerkmale: zum einen die Gebräuchlichkeit des Wortschatzes, mit tendenziell einer positiven Auswirkung auf die Textschwierigkeit, zum anderen den propositionalen Gehalt des Textes – ein Textmerkmal, das sich generell negativ auf die Textschwierigkeit auswirken soll (vgl. Abschnitt 5.2). Die Reduktion der Textschwierigkeit durch die Gebräuchlichkeit des Wortschatzes hat für die Aufgabenschwierigkeit keinen bedeutenden Gesamteffekt. Der positive Effekt auf die Aufgabenschwierigkeit kann besonders durch die Erhöhung der Textschwierigkeit durch den hohen proportionalen Gehalt gedeutet werden. Die Ergebnisse lassen sich dahingehend interpretieren, dass für diesen Faktor der proportionale Gehalt den bedeutendsten Einfluss auf die Aufgabenschwierigkeit hat.

Beschreibend: Der beschreibende Faktor hat einen geringen negativen Effekt auf die Aufgabenschwierigkeit. Die Ergebnisse können auf zwei Weisen interpretiert werden. Die erste Möglichkeit ist die Erklärung des Effektes aufgrund der Vermeidung von mathematischen Begriffen. Wie bereits für den erklärenden und instruktiven Faktor erläutert, können (mathematische) Begriffe einen Einfluss auf die Textschwierigkeit und die Aufgabenschwierigkeit zeigen. Dahingehend würde es sich eher um einen passiven negativen Effekt auf die Aufgabenschwierigkeit handeln. Die zweite Möglichkeit ist die Nutzung von textuellen Referenzenbezügen, die durch die Verwendung von direkter Anaphorik und einer hohen lexikalischen Vielfalt dargestellt sind (vgl. Abschnitt 5.2). Die Verwendung der genannten Textmerkmale kann zu einem kohärenten Text führen, der sich negativ auf die Textschwierigkeit auswirkt und damit einen reduzierenden Effekt auf die Aufgabenschwierigkeit hat. Beide Möglichkeiten können insgesamt zu einem geringen negativen Effekt auf die Aufgabenschwierigkeit führen.

Komprimierend: Der negative Effekt des komprimierenden Faktors ist dahingehend von Interesse, da in diesem Faktor Textmerkmale gemeinsam vorkommen, die nach Abschnitt 5.2 einen positiven Effekt auf die Textschwierigkeit aufweisen können. Das betrifft die Textmerkmale der durchschnittlichen Silbenanzahl, der lexikalischen Vielfalt und der Verwendung des Passivs. Der deutliche negative Effekt auf die Aufgabenschwierigkeit weist auf die Bedeutung von diskontinuierlichen Texten hin – also solchen Texten, bei denen Abbildungen zum Text verwendet werden. Durch die Nutzung von unterschiedlichen Formen von Darstellungen kann für mathematische Textaufgaben ein negativer Effekt auf die Aufgabenschwierigkeit erzeugt werden – trotz anspruchsvoller und abstrakter sprachlicher Mittel.

Die Ergebnisse machen deutlich, dass sich ein positiver bzw. negativer Effekt auf die Aufgabenschwierigkeit von Textaufgaben nicht unbedingt aus dem Einfluss auf die Textschwierigkeit von einzelnen Textmerkmalen ableiten lässt. Vielmehr muss das gemeinsame Vorkommen der Textmerkmale durch Faktoren in Beziehung gesetzt werden, um Ableitungen zum Einfluss auf die Aufgabenschwierigkeit zu treffen.

Die Ergebnisse des Modellvergleichs des LLTM und des Rasch-Modells im Hinblick auf unterschiedliche Aufgabenanforderungen: Die Ergebnisse der Analyse der Aufgabenschwierigkeiten durch die sprachlichen Faktoren machen deutlich, dass die sprachlichen Faktoren dann erklärungsstark sind, wenn die Anforderungen der Aufgabe mitbetrachtet werden. Besonders genau können die Aufgabenschwierigkeiten von Testaufgaben mit einer hohen Anforderung durch das LLTM geschätzt werden. Für leichte Testaufgaben zeigt sich eine geringe Passung der

geschätzten Aufgabenschwierigkeiten des LLTM zu den Aufgabenschwierigkeiten des Rasch-Modells. Da mit dem Instrument zur sprachlichen Variation für mathematische Textaufgaben insbesondere Textmerkmale betrachtet werden, die einen Einfluss auf die Textschwierigkeit aufweisen, lässt sich die hohe Erklärungsleistung für schwierige Testaufgaben und das häufige Überschätzen von leichten und mittelschwierigen Testaufgaben erklären. Auch aus inhaltlicher Sicht ist das Ergebnis interpretierbar. So ist für leichte und mittelschwierige Testaufgaben die Bedeutung der Textmerkmale vermutlich deshalb geringer, weil es ausreicht, das fachliche Wissen zu besitzen, um die Testaufgabe zu lösen, ohne zwangsläufig den vollständigen Aufgabentext gelesen zu haben. Um den systematischen Effekt in Bezug zur Aufgabenschwierigkeit und -leichtigkeit genauer zu erheben, ist jedoch ein Test nötig, der die Faktoren systematisch variiert, damit auch leichte Testaufgaben besser erklärt werden können. Dies war im Hinblick auf den vorhandenen Datensatz von PALMA im Rahmen dieser Arbeit aber nicht möglich. Dennoch kann das Instrument zur sprachlichen Variation leichte (durch eine negative Beziehung) und schwierige Testaufgaben genau modellieren und dafür genutzt werden, die Aufgabenschwierigkeit durch den Einbezug der sprachlichen Faktoren anzupassen.

Im Rahmen der bisherigen Erkenntnisse der Schätzung der Aufgabenschwierigkeiten durch das LLTM für das Instrument sind nun insbesondere Anpassungen für anspruchsvolle Textaufgaben denkbar. Als Ansatzpunkt können die Effekte der Faktoren auf die Aufgabenschwierigkeiten genutzt werden, um durch die aufbauende Verwendung der Faktoren die sprachlichen Anforderungen zu steigern. So hat der komprimierende Faktor den stärksten negativen Effekt auf die Aufgabenschwierigkeit, der beschreibende Faktor einen schwach negativen, der instruktive und informative haben einen moderaten positiven und der erklärende Faktor hat einen starken positiven Effekt. So kann beispielsweise der komprimierende Faktor als Ansatzpunkt dienen, fachbezogenes Vokabular rezeptiv zu erschließen, da viele typische fachliche Textmerkmale in diesem Faktor vorkommen, die jedoch aufgrund der Erleichterung durch Abbildungen vermittelt werden können. Das Ergebnis weist außerdem auf die Bedeutung von Darstellungsvernetzungen zur fachbezogenen Sprachbildung hin (Meyer & Prediger, 2012).

Die Möglichkeiten, durch die Veränderung der Faktoren die sprachlichen Anforderungen zu steigern, können sowohl im praktischen als auch im wissenschaftlichen Kontext weiter genutzt werden. In der praktischen Planung eines Mathematikunterrichts können Text-Rezipierenden-Anpassungen, durch die Faktoren des Instruments zur sprachlichen Variation von mathematischen Textaufgaben dazu beitragen, durch mathematische Textaufgaben geschaffene Lernsituationen auf die sprachlichen Voraussetzungen der Lernenden zu adaptieren. Aber

auch für Leistungssituationen bieten die Faktoren des Instruments Möglichkeiten, zu antizipieren, welche sprachlichen Hürden durch Textmerkmale im Kontext der Konstruktion eines Tests mitbedacht werden sollten.

8.5 Zusammenfassung

Das Ziel in diesem Kapitel war die Ermittlung des Effektes der Faktoren, die in Kapitel 7 für das zu konzeptualisierende Instrument extrahiert wurden, auf die Aufgabenschwierigkeit. Die Bestimmung des Effektes wurde durch die Re-Analyse des vorhandenen Datensatzes des Projektes PALMA durchgeführt. Aus diesem Grund wurden aus den gesamt betrachteten Textaufgaben diejenigen analysiert, die aus PALMA stammen. Zur Bestimmung des Effekts wurde ein LLTM durchgeführt. Im LLTM wurden die Faktoren als Basisparameter zur Bestimmung der Aufgabenschwierigkeit verwendet. Das LLTM ist eine Erweiterung des Rasch-Modells, für dessen Anwendung jedoch auch das Rasch-Modell gelten muss.

Die Ergebnisse der Rasch-Skalierung der vorhandenen Daten zeigten eine hohe Passung der Aufgaben zum Rasch-Modell. Aus diesem Grund konnte das Rasch-Modell für die vorhandenen Aufgaben angenommen und das LLTM weiterverwendet werden.

Die durch das LLTM geschätzten Aufgabenschwierigkeiten konnten eine hinreichende Passung zu den Aufgabenschwierigkeiten des Rasch-Modells erreichen, wenn alle Aufgaben betrachtet werden. In Anbetracht der Zielsetzungen konnten so die Effekte der Faktoren auf die Aufgabenschwierigkeiten ermittelt werden. Der größte positive Effekt auf die Aufgabenschwierigkeit konnte für den erklärenden Faktor ermittelt werden. Moderate positive Effekte auf die Aufgabenschwierigkeit zeigten der instruktive und informative Faktor. Negative Effekte auf die Aufgabenschwierigkeiten wiesen der beschreibende (gering) und der komprimierende (moderat) Faktor auf. Unter der Betrachtung der hinreichenden Passung der geschätzten Aufgabenschwierigkeiten des LLTM wurde zwischen Aufgabenanforderungen differenziert. Die Ergebnisse haben gezeigt, dass die Aufgabenschwierigkeit von schwierigen Testaufgaben durch die sprachlichen Faktoren deutlich genauer geschätzt werden kann als bei leichten und mittelschwierigen Testaufgaben. Dies weist darauf hin, dass die Bedeutung der sprachlichen Faktoren gegebenenfalls für diese Gruppe der Testaufgaben gering ist. Durch das Instrument können jedoch die Aufgabenschwierigkeiten von schwierigen Aufgaben erfolgreich modelliert werden.

Ausblick: Neben der zweiten quantitativen Analyse, die die Herleitungen des Effektes auf die Aufgabenschwierigkeit zum Ziel hatte, ist es ein weiteres Ziel der Konzeptualisierung des Instruments, die Faktoren mit inhaltlichen und kontextbezogenen Spezifika der mathematischen Textaufgaben in Beziehung zu setzen. Dies geschieht im dritten Teil dieser Arbeit durch eine qualitative Vertiefungsanalyse, die im nachfolgenden Kapitel dargestellt wird.