Orthografie-Trainer

Hans G. Müller

Zur Nutzung testtheoretischer Methoden als Mittel der Gestaltung individualisierter Trainingspläne im Rechtschreibunterricht – eine Dokumentation

Berlin 2011

Abstract

Der folgende Beitrag behandelt die Frage, inwieweit sich testtheoretische Methoden der empirischen Bildungsforschung nicht nur zur Analyse von Testleistungen, sondern auch zur Gestaltung von Übungseinheiten nutzen lassen. Die Überlegungen sind nicht nur theoretischer Natur, sondern haben ihren Niederschlag in der Programmierung der Internetplattform Orthografietrainer.net gefunden. Der Artikel dient somit auch der Dokumentation didaktisch relevanter Entscheidungen bei der Implementierung testtheoretischer Verfahren, ferner der Erörterung von Chancen, Problemen und Vorzügen dieser Vorgehensweise sowie der Darstellung des derzeitigen Entwicklungsstandes.

1. Aufgabenschwierigkeit als Grundlage der curricularen Trainingsgestaltung

Spätestens seit PISA ist statistisch gestützte Bildungsforschung ins Rampenlicht des breiten öffentlichen Interesses getreten. Leistungsmessung und -evaluation ist *en vouge* und das zurecht, suggerieren doch die Schulnoten seit Jahrhunderten die quantitative Messbarkeit von Lernleistungen. Dabei wurde lange Zeit überhaupt nicht infrage gestellt, ob eine solche Darstellung als numerischer Messwert überhaupt existiert oder existieren kann – und falls ja, ob sie das darstellt, was sie vorgibt: die Fähigkeiten eines Individuums in einem umgrenzten Fachbereich, sei es Physik, Mathematik – oder eben Rechtschreibung.

Durch die gängigen Auswahlverfahren an Universitäten und Berufsschulen gewinnen Leistungsmessungen einen lebenswegentscheidenden Charakter und so ist es ein didaktisches Muss, Leistungsmessungen und Leistungsskalen einer wissenschaftlichen Prüfung zu unterziehen. In zunehmendem Maße werden für diese Aufgabe auch in der Schulpraxis statistische Methoden herangezogen.

Angesichts dieser Befunde ist es durchaus verwunderlich, dass die großen statistischen Studien wie PISA lediglich im Personen-, genauer: im Ländervergleich zu so breitem öffentlichen Interesse gelangt sind. Das unerwartet schlechte Abschneiden der deutschen Schüler hat immer wieder zu Diskussionen über Schulorganisation und Schulstruktur geführt, nicht aber in gleicher Weise zur Diskussion und Überarbeitung von Lehrplänen. Dabei bieten die Messmethoden der Item Response Theorie, die bei PISA zum Einsatz kam, nicht nur Möglichkeiten, die Kompetenz von Schülern in abgrenzbaren Fachbereichen zu messen, sondern – und das ist didaktisch oft wesentlich interessanter – auch die Schwierigkeit einzelner Übungsaufgaben in diesen Fachbereichen einzuschätzen.

Ziel jedes Unterrichts ist es, die Kompetenz einer Personengruppe N im Kompetenzbereich X zu erhöhen. Der Lehrende hat dabei von den bereits vorhandenen Fähigkeiten auszugehen und ist gehalten, Strategien zu entwerfen, die systematisch zum stetigen Kompetenzzuwachs der Lerngruppe beitragen. Dabei spielt die adäquate Messung von Aufgabenschwierigkeiten eine unmittelbare, einleuchtende und didaktisch leicht umzusetzende Rolle: Was leicht ist, sollte zuerst gelernt werden, was schwerer ist, erst im Anschluss daran. Die Aufgabenschwierigkeit ist ein einfacher und plausibler Indikator dafür, welche Lernreihenfolge am erfolgversprechendsten ist. Daher ist es die Aufgabe der didaktischen Sequenzbildung, auf der Grundlage der Kenntnis von Aufgabenschwierigkeiten Lerninhalte so anzuordnen, dass sie ausgehend von der Kompetenz der Zielgruppe zur Lösung immer komplexeren Aufgaben

befähigt. Der richtigen Bestimmung von Aufgabenschwierigkeiten kommt damit eine erhebliche didaktische Bedeutung zu, da sie dazu beitragen kann, Lehrpläne zu strukturieren oder bestehende Strukturen auf ihre Angemessenheit zu prüfen.

Allerdings ist keineswegs immer unmittelbar ersichtlich, welche Aufgaben schwieriger und welche leichter sind. So gilt etwa die Kommasetzung des Deutschen als ein schwierig zu beherrschender Lerngegenstand, obwohl sie durch vergleichsweise wenige Regeln beschreibbar ist, von denen es obendrein faktisch keine Ausnahmen gibt. Gerade hier ist die Item-Response-Theorie noch zu wenig von der Didaktik beachtet worden, ermöglicht sie doch neben der Messung von Personenkompetenzen, auch die präzise Ermittlung Aufgabenschwierigkeiten. Ihr kommt damit eine ganz neue unterrichtspraktische Aufgabe zu, zu der sie bisher noch nicht in großem Rahmen herangezogen wurde: Als Strukturierungsinstrument für Lerninhalte, als Antwort auf die alltägliche didaktische Frage, in welcher Reihenfolge Unterrichtsinhalte präsentiert und erlernt werden sollten.

In der Rechtschreibdidaktik ist die Erstellung individueller Curricula von besonders hoher Bedeutung, da in kaum einem anderen Kompetenzbereich die Fähigkeiten innerhalb einer Lerngruppe so stark divergieren können. Daher ist es besonders, auf Grundlage genauer Analysen zu individualisierten Trainingsplänen zu gelangen, die den Einzelnen dort abholen, wo er gerade steht. Mithilfe statistischer Methoden lässt sich aus der Schwierigkeitsmessung von Rechtschreibaufgaben ein natürliches Rechtschreibcurriculum ableiten, das nicht von sprachwissenschaftlichen Theorien ausgeht, sondern von Fehlerarten und Fehlerhäufigkeiten, die erst in einem sekundären Prozess, quasi als Ausgangsmaterial, der sprachwissenschaftlichen Forschung zugeführt werden. Weiß man, dass die richtige Groß- und Kleinschreibung von Wort X schwieriger ist als die von Wort Y, so kann man nach sprachwissenschaftlichen Erklärungen fragen, die diesen Schwierigkeitsunterschied z. B. aus den grammatischen Eigenschaften der Wörter X und Y abzuleiten versuchen. Diese grammatischen Eigenschaften wiederum können die Grundlage der weiteren curricularen Strukturierung des Lernbereichs Rechtschreibung bilden. Bereits zuvor aber ist klar, dass in der Erwerbsreihenfolge der Groß- und Kleinschreibung Wort Y vor Wort X behandelt und geübt werden sollte.

Die lernpsychologischen Ursachen für die Schwierigkeiten von orthografischen Sachverhalten zu erforschen, hat sich Orthografietrainer.net zum mittelfristigen Ziel gesetzt. Bereits heute allerdings kann die Item Response Theorie einen erheblichen Beitrag zur Steuerung individuell angepasster Trainingspläne bieten.

Im Folgenden soll näher erläutert werden, wie dieses Vorhaben auf der Plattform Orthografietrainer.net umgesetzt wurde, welche didaktischen Entscheidungen dabei getroffen wurden und wo die derzeitigen Grenzen der Umsetzung liegen. Die vorliegende Arbeit hat damit dokumentierenden und exemplarischen Charakter. Sie ermöglicht es fachlich Interessierten einzuschätzen, was bei der Bestimmung von Satzschwierigkeiten und Schülerkompetenzen durch das Programm überhaupt geschieht. Eine Auswertung und Evaluierung des Erfolges wird durchgeführt werden, sobald genügend statistisch auswertbares Material vorliegt.

2. Item Response Theorie - Aufgabenschwierigkeiten bestimmen, Personenkompetenz messen

2.1. Das Problem traditioneller schulischer Testmethoden

Will man eine bestimmte Menge an Aufgaben ihrer Schwierigkeit nach ordnen, um daraus eine natürliche Erwerbsreihenfolge abzuleiten, so kommt der richtigen Messung von Aufgabenschwierigkeiten eine entscheidende Rolle zu. In der alltäglichen Schulpraxis hingegen spielt meist lediglich die Messung von Personenkompetenzen eine Rolle, die durch Aufsummieren von richtig und falsch gelösten Aufgaben erfolgt. Aus dem Prozentsatz erreichter Punkte wird nach festgelegten Schlüsseln die Note errechnet, die die Kompetenz eines Schülers wiedergeben soll.

Die Literatur zur Kritik an der klassischen Notengebung in der Schule ist umfangreich und soll hier nicht referiert werden (vgl. dazu Jürgens 2005: 44ff). Kritiker bemängeln zurecht die Schein-Objektivität dieses Vorgehens, Verteidiger führen zurecht ins Feld, dass es derzeit wenige praktikable Alternativen gibt. Für die hier interessierende Fragestellung ist hingegen hervorzuheben, dass die Summierung von Punkten implizit von der Annahme ausgeht, dass jede Aufgabe gleich schwierig ist, da ihre Lösung den gleichen Beitrag zur Kompetenzbestimmung leistet wie alle anderen. Zwar können Lehrer für eine Aufgabe mehrere Punkte verteilen und damit eine Art Gewichtung vornehmen, doch basiert auch diese Gewichtung nicht auf Messungen, sondern auf einer mehr oder weniger theoretisch fundierten didaktischen bzw. lehrpraktischen Intuition des Lehrenden.

Ferner wird mit der Summierung von Punkten implizit die Behauptung aufgestellt, der Test oder die Aufgaben würde eine und nur eine bestimmte Kompetenz messen. Ob dabei nicht eigentlich Äpfel und Birnen addiert werden, kann nicht geprüft werden, ja es wird zumeist noch nicht einmal als Problem erkannt.

Längst weiß die Rechtschreibdidaktik, dass Rechtschreibfehler nicht gleichwertig sind, und kommt zunehmend von der Tradition ab, Fehler zu summieren und Noten auf der Grundlage von Fehlerquotienten zu geben. Gängige standardisierte Testverfahren der Rechtschreibkompetenz, so etwa die HSP (May et al. 2002) oder der DRT (Grund et al. 2004), nehmen qualitative Fehlerkategorisierungen vor, um damit mehrere Teilkompetenzen zu unterscheiden. Die Einordnung der Fehler orientiert sich dabei an orthografischen Theorien oder Schriftspracherwerbsmodellen (für einen Überblick vgl. jüngst Fay 2010: 43ff). Innerhalb der postulierten Fehlerkategorien erfolgt in der Regel wiederum eine quantitative Messung durch Fehleraddition. Obwohl die derzeit verfügbaren standardisierten Rechtschreibtests damit die Möglichkeiten der Testtheorie keineswegs ausschöpfen, haben sie gegenüber dem einfachen Zählen von Rechtschreibfehlern entscheidende Vorteile, zeigen sie doch genauer, in welchen Bereichen didaktische Intervention von besonderer Bedeutung ist. Zwei Probleme allerdings lassen sich mit dieser Vorgehensweise nicht einschätzen: Erstens die Frage, ob die angenommenen Fehlerkategorien die mentalen Prozesse des Schreibens bestmöglich modellieren, und zweitens, welche Erwerbsreihenfolge orthografischer Phänomene die größten Lernerfolge erzielen kann. Beide Fragen sind oft und intensiv theoretisch untersucht, kaum aber systematisch statistisch aufgearbeitet worden.
In der Praxis des Unterrichtsalltags schließlich halten sich die traditionellen Verfahren der Berechnung einfacher Fehlerquotienten erstaunlich hartnäckig, wahrscheinlich, weil derzeit noch zu selten praktikable Alternativen für den Schulalltag angeboten werden.[1]

2.2. Das Rasch-Modell und die Vorteile testtheoretischer Methoden

Probleme der Leistungsmessung haben in der empirischen Bildungsforschung zu einer ganzen Reihe theoretischer Ansätze geführt, die unter dem Namen Item Response Theorie (IRT) zusammengefasst werden. (zum Begriff in Abgrenzung von der „klassischen“ Testtheorie vgl. Rost, 2004: 12). Sie alle basieren auf der gemeinsamen Überlegung, Aufgabenschwierigkeit und Personenkompetenz auf ein und der selben Skala abzubilden. Eine Aufgabe ist offenbar umso schwieriger, je weniger Personen sie lösen können, und umso leichter, je mehr Personen dazu imstande sind. Die Kompetenz einer Person wiederum lässt sich entsprechend an Art und Menge der von ihr gelösten Aufgaben ablesen. Es werden folglich Personenkompetenz

[1] So ist etwa im Land Berlin seit 2009 eine kriteriumsorientierte Bewertung der Rechtschreibleistung vorgeschrieben (vgl. Verwaltungsvorschrift Schule 3/2009). Diese versucht zwar, die Bewertung der Rechtschreibung an Kriterien festzumachen, die für die schriftliche Kommunikation tatsächlich relevanter sind als die bloße Fehlerhäufigkeit, nämlich der Beeinträchtigung der Verständlichkeit und des Leseflusses. Die fehlende Operationalisierbarkeit dieser Kriterien führt allerdings nicht zu einer größeren Sicherheit in der Leistungsbewertung, sondern kann im Gegenteil sehr willkürlich ausgelegt werden.

und Aufgabenschwierigkeit gegenseitig aneinander gemessen.

Bis dahin unterscheidet sich die Item Response Theorie gar nicht weit von der Vorgehensweise in der Schule, da in beiden Fällen Lösungshäufigkeiten summiert werden, um eine Rangfolge zu bilden. Der entscheidende Unterschied ist allerdings der, dass es in der IRT umfassende Prüfmethoden gibt, ob eine solche Summierung zulässig ist oder nicht. Das ist nämlich genau dann der Fall, wenn (unter Berücksichtigung statistischer Schwankungen) auch nur die kompetentesten Personen die schwierigsten Aufgaben lösen konnten, während Personen niedriger Kompetenz die leichtesten, nicht aber die schwereren Aufgaben lösen können sollten. Abb. 1 stellt das Problem dar: Während in 1a) nur die kompetentesten die Personen auch die schwierigsten Aufgaben lösen können und sich somit eine geordnete Matrix von Lösungen und Nichtlösungen erstellen lässt, ist dies in 1b) offenbar nicht in gleicher Weise der Fall.

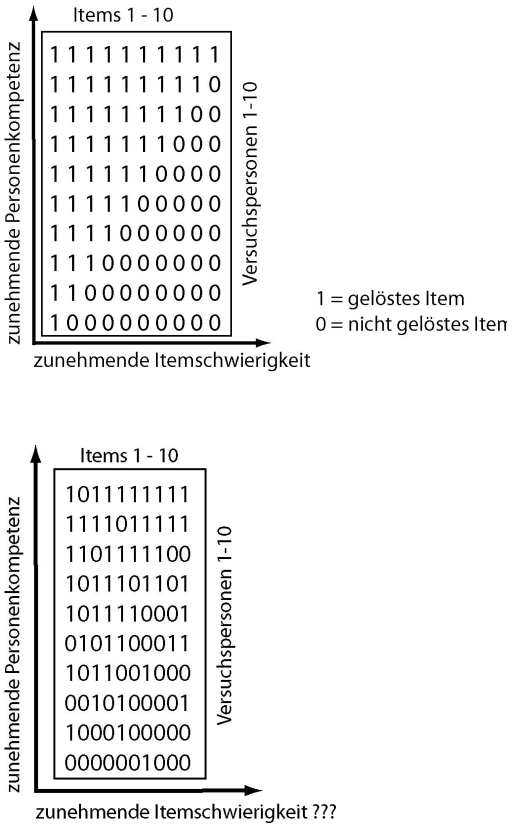

Abb. 1: divergierende Lösungsmatrizen

Stellt sich in einem Test heraus, dass sich Personenkompetenzen und Aufgabenschwierigkeiten nicht (annähernd) wie in 1a) ordnen lassen, so stimmt etwas nicht an der zugrunde gelegten Theorie: Sie misst dann offenbar nicht eine Kompetenz, sondern bspw. eine Mischung aus zwei oder mehr Persönlichkeitsvariablen. Eine Summierung ist in diesem Falle nicht zulässig.

Der Zusammenhang von Aufgabenschwierigkeit und Personenkompetenz wird je nach zugrunde gelegter Theorie durch eine bestimmte Modellgleichung hergestellt. Dabei kommt dem von dem dänischen Mathematiker Georg Rasch (1960) entwickelten und nach ihm benannten Modell eine besondere Bedeutung zu. Das Rasch-Modell verfolgt einen probabilistischen Ansatz, d.h. es stellt eine Beziehung zwischen der Personenkompetenz und der Lösungswahrscheinlichkeit einer bestimmten Aufgabe her. Abb. 2 zeigt eine solche Beziehung anhand einer Testaufgabe: Mit zunehmender Kompetenz der Testperson nimmt die Lösungswahrscheinlichkeit für diese Aufgabe stetig zu.

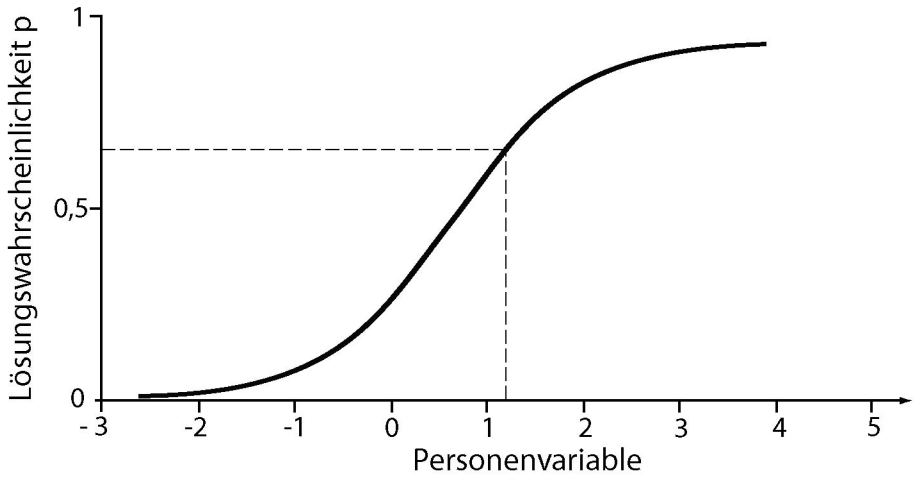

Abb. 2: Item caracteristic curve (ICC) aus Personenkompetenz und Lösungswahrscheinlichkeit

Der ogivenförmige Verlauf der Kurve ist dabei einerseits durch mathematische Überlegungen bedingt, auf die hier nicht weiter eingegangen werden soll (vgl. dazu Rost, 2004: 115), ist aber andererseits auch lernpsychologisch plausibel, da sich bei extrem niedrigen und extrem hohen Kompetenzen die Lösungswahrscheinlichkeit für eine Aufgabe beinahe gar nicht ändert, während um den Wendepunkt der Kurve herum eine nahezu lineare Zunahme der Lösungswahrscheinlichkeit zu verzeichnen ist: Schulpraktisch ausgedrückt heißt das: Hat ein Schüler eine bestimmte Kompetenz θ erreicht, so steigt die Lösungswahrscheinlichkeit für eine Aufgabe i der Schwierigkeit σ kontinuierlich an. Liegt die Kompetenz des Schülers weit unterhalb oder weit oberhalb der Schwierigkeit σ, so ist die Lösungswahrscheinlichkeit für i nahezu 0 bzw. nahezu 100%.

Große internationale Studien wie PISA oder TIMSS nutzen das Rasch-Modell zur Bestimmung ihrer Messwerte: Im Test wird die Aufgabenschwierigkeit jeder Testaufgabe anhand der Modellgleichung bestimmt. Anschließend erhält jeder Schüler denjenigen Wert als Kompetenzwert zugeschrieben, an dem seine Lösungswahrscheinlichkeit bei einem festgelegten Wert liegt (oft wird eine Wahrscheinlichkeit von 65% gewählt, vgl. Gonzalez 1997: 150). Für die in Abb. 2 dargestellte Testaufgabe wäre dies etwa bei einem Messwert von 1,2 der Fall (vgl. gestrichelte Linie).

2.3. Die pair-wise Methode

Im letzen Abschnitt ist deutlich geworden, dass die Berechnung von Aufgabenschwierigkeit und Personenkompetenz mithilfe des Rasch-Modells wichtige Vorteile gegenüber dem Auszählen von Fehlerhäufigkeiten hat. Leider stehen dem Einsatz des Rasch-Modells im didaktischen Alltag zwei Umstände im Wege: Zum einen benötigt man für die angemessene

Schätzung der Parameter im Rasch-Modell vergleichsweise große Datenmengen, zum anderen lässt sich die Modellgleichung des Raschmodells nicht einfach nach einer bestimmten Variable hin auflösen – sprich: Aufgabenschwierigkeit und Personenkompetenz lassen sich nicht einfach *berechnen*, sondern müssen durch rechenintensive iterative Verfahren ihrem wahrscheinlichsten Wert angenähert werden.

Die erste Einschränkung stellt für Orthografietrainer.net keine größeres Hürde dar. Im Gegenteil kann das Portal hier seine entscheidenden Stärken ausspielen und die Parameterschätzung auf eine breite Datengrundlage von mehreren Zehntausend Lösungen pro Testsatz stützen.

Iterative Schätzverfahren hingegen sind im Online-Bereich problematisch, da sie erhebliche Rechenlast verursachen und so zur Laufzeit des Servers nicht ohne Weiteres durchgeführt werden können. Ferner bedingen sie bereits im Vorfeld großen Programmieraufwand.

Auch dieses Problem wird in späteren Programmversionen gelöst werden können. Bis dahin gilt es, Möglichkeiten zu finden, das Potenzial einer großen Datenbasis didaktisch nutzbar zu machen. Für die Aufgabenschwierigkeit existiert eine explizite Berechnungsformel, welche Werte liefert, die denen des Rasch-Modells äquivalent sind, die sog. pair-wise-Methode (vgl. Rost 2004: 310). Dieser Ansatz berechnet das Schwierigkeitsverhältnis zwischen einer Testaufgabe i und einer anderen Testaufgabe j daraus, wie oft i gelöst werden konnte, j hingegen nicht (n_{ji}), bzw. im umgekehrten Fall, wie oft j gelöst wurde, nicht aber i (n_{ij}). Diese beiden Werte stehen in direkter Proportion zum Exponenten der Aufgabenschwierigkeit von j und i, sodass sich zur Berechnung der Aufgabenschwierigkeit einer Testaufgaben in Relation zu allen anderen Testaufgaben die folgende Berechnungsformel ergibt:

$$\sigma_j = \left(\frac{1}{k}\right) \sum_{i \neq j} \left(\log\left(n_{ij}\right) - \log\left(n_{ji}\right)\right)$$

Abb. 3: Pair-wise-Algorithmus

Nach dieser Formel berechnet Orthografietrainer.net derzeit die Aufgabenschwierigkeiten.

2.4. Vorteile und Grenzen der pair-wise-Methode

Der pair-wise-Algorithmus ermöglicht es, zur Laufzeit des Programms die Aufgabenschwierigkeiten zu berechnen und aktuell zu halten. Mit zunehmender Lösungshäufigkeit eines Testsatzes stützt sich die Berechnung seiner Schwierigkeit auf eine

immer größere Datenmenge. Die berechneten Aufgabenschwierigkeiten sind denen des Rasch-Modells äquivalent (vgl. Rost 2004: 311) und verfügen über deren Vorzüge im Gegensatz zur bloßen Auszählung von Fehlerhäufigkeiten. Darüber hinaus verbindet sich mit dem pair-wise-Algorithmus ein für das Testdesign von Orthografietrainer.net wichtiger Vorteil: der unproblematische Umgang mit unvollständigen Datenmatrizen. Kaum ein Benutzer von Orthografietrainer.net löst systematisch jeden Übungssatz. Daraus folgt, dass die Daten, die das Programm im Bezug auf sein gesamtes Übungsrepertoire speichert, systematische Lücken aufweisen. Die pair-wise-Methode bezieht nur diejenigen Teile der Datenmatrix in die aktuelle Berechnung ein, in der zu je zwei Aufgaben i und j auch Daten ein und derselben Person vorliegen. Damit erledigt sich nebenbei auch das Problem, dass die Aufgabenlösungen auf Orthografietrainer.net nicht zufällig auf bestimmte Nutzergruppen verteilt sind, sondern systematische Häufungen aufweisen: Da schon durch Lehrplanvorgaben bestimmte Bereiche der Rechtschreibung für bestimmte Klassenstufen typisch sind, ist es kein Zufall, dass gerade leichte Aufgaben von jüngeren Nutzern wesentlich häufiger bearbeitet werden, während ältere Nutzer tendenziell komplexere Aufgaben lösen. Für den pair-wise-Algorithmus ist dieser Umstand unproblematisch, da für die Schwierigkeitsrelation zwischen den Aufgaben i und j lediglich die Daten derjenigen Probanden herangezogen werden, die auch beide Aufgaben gelöst haben.

Neben diesen Vorteilen sind zwei Grenzen zu erwähnen, denen im Folgenden jeweils ein eigenes Kapitel gewidmet ist: Erstens kann die Gültigkeit des Rasch-Modells für die Daten von Orthografietrainer.net zur Laufzeit nicht geprüft werden, womit derzeit ein Vorteil der IRT noch nicht zum Tragen kommen kann. Bis zu einer ausführlichen statistischen Studie muss daher die Rasch-Skalierbarkeit der Daten eine Hypothese bleiben (die Konsequenzen dieses Umstandes werden in den folgenden Kap. 3.1 und 3.2 ausführlich diskutiert). Zweitens liefert die pair-wise-Methode keine Messwerte für die Personenkompetenz, sodass hier auf andere Schätzungen ausgewichen werden muss, die in Kap. 4.3 erläutert werden. Wie in diesen Kapiteln eingehend dargestellt werden wird, gibt es dennoch gute Gründe, die Schätzwerte der pair-wise-Methode zur Erstellung individualisierter Trainingspläne heranzuziehen, da sie das beste derzeit praktikable Verfahren darstellt.

3. Die Hypothese der Rasch-Skalierbarkeit – Vorstellung und Kritik

3.1. Wie viele Rechtschreibkompetenzen gibt es?

Im letzten Kapitel war angesprochen worden, dass ein wichtiger Vorteil des Rasch-Modells darin besteht, dass geprüft werden kann, ob die Aufgaben eines Tests eine und nur eine Personenvariable ansprechen oder ob diese Annahme aufgegeben werden muss. Solche Modellgeltungstests gibt es in großer Zahl und sie haben verschiedene Vor- und Nachteile (vgl. Rost 2004: 330ff). Tests, die auf die Datenmatrizen von Orthografietrainer.net anwendbar wären (z. B. das sog. Bootstrap-Verfahren, Rost 2004: 337f, Langeheine, Pannekoek, van de Pol 1996: 496ff) sind so rechenintensiv, dass sie zur Laufzeit der Plattform die Rechenkapazität überfordern würden. Hingegen können sie bei der wissenschaftlichen Auswertung der Daten von Orthografietrainer.net lokal berechnet werden. Bis dies geschehen kann, stellt sich aber die Frage, ob die online verfügbaren Messwerte nach der pair-wise-Methode bereits eine praktikable Grundlage bilden, auf der individualisierte Trainingspläne für jeden Übenden erstellt werden können. Letztlich wird sich auch diese Frage mit zunehmender Nutzung des Angebots von Orthografietrainer.net statistisch beantworten lassen. Bis dahin gilt es allerdings, die theoretischen Überlegungen zu dokumentieren, die die Hypothese der Rasch-Skalierbarkeit stützen. Außerdem muss überlegt werden, welche Konsequenzen zu erwarten sind, wenn sich die Hypothese als nicht haltbar erweist.

Anhand des aktuellen Forschungsstandes zur Orthografie ist nicht abschließend zu beantworten, ob es sich bei der Rechtschreibkompetenz um eine oder mehrere Personenvariablen handelt. Denkbar wäre, dass es eine einzige Personenfähigkeit gibt, die die Rechtschreibleistung dieser Person adäquat abbildet. Denkbar wäre genauso, dass eine Person für jeden Bereich der Rechtschreibung (Laut-Buchstaben-Zuordnung, Getrennt- und Zusammenschreibung, Groß- und Kleinschreibung, Kommasetzung) unterschiedliche Kompetenzen an den Tag legt. Schließlich ist auch die Hypothese möglich, dass selbst die einzelnen Bereiche der Rechtschreibung in unterschiedliche Kompetenzbereiche zerfallen – bspw. in eine Kompetenz für die Kommatierung von Gliedsätzen und eine für die Kommatierung von Appositionen etc. Diese Teilkompetenzen könnten parallel zu den einzelnen Rechtschreibregeln liegen, aber genau so gut eine ganz andere Struktur aufweisen.

Angesichts dieser theoretischen Ausgangslage stellt sich die Frage, welche und wie viele Kompetenzen Orthografietrainer.net annehmen soll, um die Schwierigkeiten jeder Aufgabe bzw. die Fähigkeiten jeder Person adäquat abzubilden.

Die Passung von Testmodellen auf statistisches Datenmaterial ist nie eine absolute Entscheidung. Die wissenschaftliche Praxis hat sich auf bestimmte Kriterien geeinigt, unter denen die Geltung eines Testmodells aufrechterhalten oder zurückgewiesen werden sollte. Darüber hinaus werden solche Kriterien auch für die Frage verwendet, ob ein Modell die Daten besser erklären kann als ein anderes. Insbesondere bei dieser Frage spielen aber auch Faktoren eine Rolle, die nicht eigens berechnet werden müssen, so etwa das Kriterium der Einfachheit: Bei ähnlicher Modellpassung oder dem Fehlen von Modellgeltungstests ist etwa das einfachere Modell dem komplexeren vorzuziehen, was wissenschaftstheoretische wie statistische Gründe hat (so etwa die Menge an zu schätzenden Parametern, vgl. Rost 2004: 263). Die einfachste Annahme über den Lernbereich Rechtschreibung ist, dass es eine und genau eine Rechtschreibkompetenz gibt, die jede Person in unterschiedlichem Maße aufweist. Seine Kompetenz entspricht damit in etwa dem Prozentsatz richtig gelöster Aufgaben. Von dieser Annahme geht implizit jeder aus, der verschiedene Rechtschreibfehler addiert und daraus Fehlerquotienten berechnet.

Um es Lehrern wie Schülern zu ermöglichen, die Online-Leistungen direkt mit den in der Schule erhobenen Fähigkeiten zu vergleichen, stellt Orthografietrainer.net den Prozentsatz richtig gelöster Aufgaben als Messwert zur Verfügung. Darüber hinaus berechnet das Programm mithilfe der pair-wise-Methode eine Skala der Gesamtschwierigkeit bzw. Gesamtkompetenz, die gewissermaßen das testtheoretische Pendant zu diesem Prozentwert darstellt.

Die Annahme, bei der Rechtschreibleistung handle es sich um genau eine Kompetenz, ist in vielerlei Hinsicht unbefriedigend und vermutlich inadäquat. Zwar kann sie, wie gesagt, auf eine lange Tradition zurückblicken und verbucht das Kriterium der Einfachheit für sich, doch bereits die schulpraktische Erfahrung deutet auf die Existenz mehrerer, voneinander mehr oder weniger unabhängiger Kompetenzen hin: Jeder Lehrer, der Fehlerkarteien mit unterschiedlichen Abteilungen zu einzelnen Bereichen der Rechtschreibung führen lässt, vertritt damit implizit die Annahme mehrerer unterschiedlicher Kompetenzbereiche. Auch im Alltag finden sich Belege für die Annahme unterschiedlicher Kompetenzen: Ein Schreibender kann von sich behaupten, die Groß- und Kleinschreibung zu beherrschen, aber in der Kommasetzung Probleme zu haben, was testtheoretisch betrachtet der Annahme unterschiedlicher Personenkompetenzen entspricht.

Schließlich machen auch die Herkunft und Festlegung der Rechtschreibregeln selbst die Annahme mehrerer Kompetenzen plausibel: So liegen der Laut-Buchstaben-Zuordnung vorrangig phonologische und morphologische Kriterien zugrunde, der Groß- und

Kleinschreibung morphologische und syntaktische, der Kommasetzung vorrangig syntaktische. Warum sollten Regeln mit so unterschiedlichem normativen Fundament einen einzigen Kompetenzbereich bilden?

Aus diesen Gründen berechnet Orthografietrainer.net neben einer Gesamtskala der Rechtschreibkompetenz außerdem für jeden Teilbereich der Rechtschreibung eine Einzelskala nach der pair-wise-Methode. Auch für diese Teilbereiche stehen zusätzlich Prozentwerte der Fehlerhäufigkeit zur Verfügung. Derzeit ist nicht geklärt, ob die Gesamtskala, die vier Einzelskalen oder vielleicht ein ganz anderes Modell die Rechtschreibdaten am besten erklärt. Nach den oben genannten Kriterien und theoretischen Überlegungen gibt es allerdings triftige Gründe, die Rasch-äquivalenten Messwerte als vorläufig beste Schätzer der Aufgabenschwierigkeit heranzuziehen.

Obwohl noch kein umfassendes Bild der orthografischen Kompetenz vorliegt, kann für Teilbereiche der Rechtschreibung bereits jetzt die Rasch-Skalierbarkeit bestätigt werden. So konnte der Autor dieses Artikels 2006 die Rasch-Skalierbarkeit im Bereich der Kommasetzung belegen (Müller 2007: 113f). In der gleichen Arbeit gab es auch (noch nicht hinreichend gesicherte) Anhaltspunkte dafür, dass die postulierte Kommakompetenz möglicherweise in mehrere Teilkompetenzen zerfallen könnte (ebd.: 152ff). Bezeichnenderweise zeigten diese Hinweise, dass sich die möglichen Teilkompetenzen nicht mit den einzelnen Kommaregeln deckten, sondern eher parallel zu den Teilbereichen der Sprachwissenschaft: Phonologie, Syntax und Semantik lagen – und das, obwohl bspw. die Phonologie für die Regelung der deutschen Kommasetzung keine Rolle spielt. Diese Befunde sind für den hier interessierenden Zusammenhang insofern von Bedeutung, als sie erstens einen Beleg für die Rasch-Skalierbarkeit der orthografischen Teilbereiche liefern, gleichzeitig aber zweitens Hinweise darauf geben, dass die bisher vorgenommene Skalierung zwar ein sinnvolles, aber möglicherweise nicht das beste Modell der Rechtschreibkompetenz bilden und dass sich drittens bessere Modelle nicht notwendigerweise durch die Orientierung an etablierten orthografischen Theorien ergeben.

3.2. Konkurrierende Hypothesen und ihre Auswirkungen

Mehrdimensionale Raschhypothesen zu prüfen, erfordert eine erhebliche Menge an statistischen Daten, da die Anzahl zu schätzender Modellparameter im Gegensatz zum einfachen Rasch-Modell stark zunimmt (vgl. Rost 2004: 263). Die Tatsache, dass mehrere vielversprechende Erhebungen auf der Basis mehrdimensionaler Modelle keine bessere Modellpassung nachweisen konnten als die eindimensionalen Modelle, kann durchaus von

diesem Problem mitverursacht worden sein (vgl. etwa Rupp 2009: 117f, Kunina-Habenicht, Rupp, Wilhelm 2009: 64). Bevor also eine Prüfung komplexerer als der bisher angenommenen Testmodelle überhaupt erfolgversprechend sein kann, müssen zunächst hinreichend große Datenmengen erhoben worden sein.

Solange die Rasch-Skalierbarkeit der Rechtschreibkompetenz bzw. der einzelnen Teilbereiche nicht geprüft und bestätigt wurde, gilt es, zu überlegen, was der mögliche Fall des Scheiterns der Rasch-Hypothese für die bis dahin verwendeten Messwerte bedeutet. Schließlich werden aufgrund dieser Messungen Kompetenzen geschätzt und Aufgabenempfehlungen gegeben. Ein solches Vorgehen ist nur dann unbedenklich, wenn die Empfehlungen selbst im Falle der Nicht-Skalierbarkeit noch angemessen didaktisch interpretierbar sind.

Dass dies der Fall ist, erhellt sich aus der Tatsache, dass die Messdaten selbst im ungünstigsten Fall nicht mehr, aber auch nicht weniger aussagen würden als die einfachen Auszählung der Fehlerhäufigkeiten. Gegenüber diesen hätten sie hingegen immer noch den Vorteil, das Problem des systematisch unterschiedlichen Alters der Versuchspersonen einzubeziehen. Die größte Gefahr in der Nutzung der berechneten Aufgabenschwierigkeiten liegt folglich darin, *nicht besser* zu sein als die gängige schulische Praxis. Unter diesen Bedingungen kann eine Nutzung der vorläufigen Daten für unbedenklich erklärt werden.

3.3. Fragen der Auflösung: Sätze oder Wörter?

Das Übungsdesign von Orthografietrainer basiert auf der Präsentation ganzer Sätze, die ein oder mehrere orthografische Probleme enthalten. Dieser Umstand liegt zum einen in der didaktischen Überlegung begründet, orthografische Probleme in ein möglichst typisches semantisches Umfeld einzubetten und damit die Ähnlichkeit zwischen Übungs- und Schreibsituation zu erhöhen. Zum anderen lassen sich einige Teilbereiche der Rechtschreibung – so etwa die Kommasetzung – gar nicht anders als im Bezug auf den Satz bearbeiten. Außerdem ist es für den Lernprozess wünschenswert, wenn die Schüler nicht im Voraus wissen, wie viele Probleme in einem Übungssatz auf sie warten, um ihre Aufmerksamkeit nicht im Voraus künstlich einzuschränken.

Im Gegensatz dazu wäre es aus testtheoretischer Sicht wünschenswert, Testaufgaben einzeln zu betrachten, sie also möglichst auch einzeln zu präsentieren. Das ist im orthografischen Bereich aus den oben genannten Gründen nicht immer möglich.

Die Messung der Aufgabenschwierigkeit basiert derzeit auf dem Übungssatz als Ganzem, selbst wenn dieser eigentlich aus mehreren orthografischen Teilproblemen besteht. Die Schwierigkeit eines solchen komplexen Rechtschreibproblems lässt sich nicht ohne Weiteres

als die Summe der Teilprobleme begreifen, ist aber auch nicht einfach mit dem schwierigsten Einzelproblem des Satzes gleichzusetzen. Daraus ergibt sich ein Interpretationsproblem, denn das Zustandekommen der Aufgabenschwierigkeit eines Testsatzes lässt sich somit testtheoretisch schwer fassen. Für den Zweck der Anordnung der Testsätze im Trainingsplan hingegen ist dieser Mangel unerheblich, da hier die bloße Tatsache hinreicht, dass der Satz bei einer Kompetenz θ mit der Wahrscheinlichkeit p gelöst wird. Welche Prozesse dabei im Lösenden ablaufen, ist vorerst vernachlässigbar. Gleichwohl steckt in diesem Problem Optimierungspotenzial für die weitere Arbeit, da das Übungsmaterial umso effizienter und konsistenter wird, je klarer ein Aufgabensatz die Fortentwicklung der Personenkompetenz in einer ganz bestimmten Situation des Rechtschreiberwerbs unterstützt. Dieser Umstand deutet für die Zukunft der Entwicklung von Orthografietrainer.net ein rekursives Vorgehen an: Die genaue, auf Einzelitems basierende Analyse der Ursachen für Rechtschreibschwierigkeiten werden Schritt für Schritt zur Umstrukturierung des Test- und Übungsmaterials führen, die ihrerseits wiederum neue Messungen notwendig machen.

4. Zur Erstellung individualisierter Trainingspläne - Vorgehensweise von Orthografietrainer.net

4.1. Aufgabenauswahl – qualitative Messung

Sinn und Zweck der Bestimmung von Aufgabenschwierigkeit und Personenkompetenz ist es, jedem Schüler diejenigen Aufgaben zur Lösung anzubieten, die einerseits seinem Fehlerprofil entsprechen und andererseits seine derzeitige Kompetenz nicht in einem Maße übersteigen, dass Frustrationen vorprogrammiert wären.

Die meisten Übungsaufgaben von Orthografietrainer.net fokussieren einen bestimmten auf den Phänomenen der deutschen Orthografie beruhenden Regelkomplex. Da diese Regeln auch Grundlage der Testsätze sind, scheint eine Zuordnung von Fehlern im Test und daraus resultierender Aufgabenauswahl zunächst vergleichsweise unproblematisch: Sie basieren auf der Hypothese, dass ein Fehler im Regelkomplex X ein Hinweis auf mangelnde Kompetenz in diesem Komplex darstellt. Folglich werden Übungssätze ausgewählt, die eben diesen Regelkomplex behandeln.

Allerdings muss es durchaus nicht sein, dass sich individuelle Fähigkeiten und Probleme eines Übenden am besten durch die Struktur der orthografischen Theorie erklären lassen, denn die mentalen Repräsentationen orthografischen Wissens können ganz anders strukturiert sein

als die bestehenden Regeln der deutschen Rechtschreibung. In Kap. 3.1 wurde bereits exemplarisch diskutiert, dass das theoretisch Einfache nicht notwendig auch einfachen praktischen Erwerb bedingt. Für den Bereich der Kommasetzung konnte z. B. nachgewiesen werden, dass semantische und phonologische Kriterien für die Schwierigkeit einer Kommastelle eine wichtigere Rolle spielen als syntaktische, obwohl die entsprechenden Regeln ausschließlich auf syntaktischen Kategorien aufbauen (vgl. Müller 2007: 129). Auch in der Groß- und Kleinschreibung zeichnen sich bereits Hinweise für einen solchen Unterschied von externen Regeln mentalen Repräsentationen ab, die allerdings noch der genaueren Untersuchung bedürfen.

Andererseits bietet die Regelstruktur vorerst einen gangbaren und plausiblen Weg, Rechtschreibfehler qualitativ einzuordnen. Mit zunehmender Kenntnisse über die Struktur des mentalen Rechtschreibwissens kann bzw. muss die Orientierung an den Rechtschreibregeln ggf. aufgegeben werden.

Die qualitative Einordnung von Rechtschreibfehlern durch Orthografietrainer.net geschieht daher auf doppelte Weise. Einerseits spiegeln die Testsätze eine interne Systematik wider, die auf statistischen Voruntersuchungen der Datenbasis von Orthografietrainer.net beruht und zum Teil den etablierten qualitativen Analyserastern wie HSP oder DRT folgt, zum Teil auch andere, bisher nicht systematisch einbezogene Kriterien berücksichtigt, die sich in der Voruntersuchung als signifikante Einflussfaktoren erwiesen haben. Andererseits geschieht eine Zuordnung zu entsprechenden Rechtschreibübungen nach den in der jeweiligen Übung fokussierten Phänomenen (vgl. dazu die Übungsauswahl von Orthografietrainer.net). Mit zunehmender Kenntnis der grammatischen Einflussfaktoren auf die Schwierigkeit eines orthografischen Problems wird die Zuordnung von Übungssätzen zunehmend von diesen Einflussfaktoren bestimmt werden.

Im Bereich der Laut-Buchstaben-Zuordnung werden derzeit fünf qualitative Kriterien gesondert berücksichtigt und in den Testsätzen systematisch variiert, die den Voruntersuchungen gemäß signifikanten Einfluss auf die Schwierigkeit einer Wortschreibung haben. Die Kriterien beziehen sich auf jeweilige Abweichungen vom jeweiligen „Normalfall“ der deutschen Rechtschreibung, etwa in den Regeln der Phonem-Graphem-Korrespondenzen oder der silbischen Struktur sowie nämlich:

- Besonderheiten in der Konsonantenschreibung (etwa <v> statt <f>)
- Besonderheiten in der Vokalschreibung (etwa Doppelvokale)
- Besondere Länge- oder Kürzemarkierungen (z.B. Dehnungs-h)

- Irregularitäten im Silbenaufbau (z.B. Doppelkonsonanten vor nicht-betonten kurzen Vokalen)
- Morphologisch bedingte Besonderheiten (z. B. Auslautverhärtung)

Die Getrennt- und Zusammenschreibung zeigte in der Voruntersuchung bezüglich der Aufgabenschwierigkeit viele Ähnlichkeiten, aber auch starke Abweichungen von den gängigen Regeln der deutschen Orthografie. Einerseits scheinen, wie in der Systematik der amtlichen Regelungen die Wortartenkategorien der infrage stehenden Glieder eine Rolle zu spielen, andererseits zeigten sich auch semantische und phonologische Kriterien als signifikant. In die systematische Variation von Einflussfaktoren wurden vier Kriterien einbezogen:

- Neue Bedeutung der Verbindung im Gegensatz zur getrennten Schreibung
- Wortart des Erstglieds
- Wortart des Zweitglieds
- Betonungsstruktur von Erst- und Zweitglied

Bei der Groß- und Kleinschreibung erwies sich eine so große Fülle von möglichen Einflussfaktoren als signifikant, dass nur ein Teil der möglichen Einflussfaktoren innerhalb der Testsätze systematisch variiert werden konnte. Gleichszeitig zeigte sich eine besonders starke Abweichung von den amtlichen Regelungen, die Groß- und Kleinschreibung nach wie vor an die Wortartenkategorie binden. Eine bedeutendere Rolle scheinen, wie auch in der didaktischen Forschung diskutiert syntaktische Einflussfaktoren zu haben(vgl. RÖBER-SIEKMEYER 1999, GÜNTHER, NÜNKE 2005). Insgesamt wurden die folgenden Kriterien berücksichtigt:

- Semantik des Wortes (Konkretum/Abstraktum)
- Vorhandensein eines Artikels/einer Mengenangabe
- Abhängigkeit von einer Präposition
- Derivationsmerkmale (typische Substantivendungen, Desubstantivierungen etc.)

Die Kommasetzung orientiert sich an den vom Autor zu diesem Gebiet erbrachten Forschungen (MÜLLER 2007), wobei designbedingt phonologische Einflussfaktoren (Sprechpausen) in den Tests und Übungen nicht kontrolliert werden können. Folgende Kriterien sind in die Entwicklung der Testsätze eingegangen:

- Satzhaftigkeit (nicht verbal / infinit-verbal / finit)
- Syntaktische Hierarchie (Über-, Unter- oder Nebenordnung)
- Semantische Funktion der syntaktischen Struktur
- Vorhandensein von Signalwörtern

Ob und inwieweit die vorgestellten Kategorien sich auch in der Hauptuntersuchung als adäquat und hinreichend differenziert für die Einordnung der Rechtschreibprobleme erweisen werden, ist späteren Untersuchungen vorbehalten.

4.2. Aufgabenauswahl – quantitative Messung

Während die qualitative Bestimmung der richtigen Übungen für einen Schüler vorerst keine entscheidenden Probleme bereiten, ist es notwendig, anhand der quantitativen Messwerte eine Entscheidung darüber treffen zu können, ob die anhand des Fehlerprofils ausgewählten Übungen auch mit hinreichendem Lernerfolg bewältigt werden können. Eine Aufgabe sollte so schwer sein, dass der Übende sie zwar nicht fehlerlos bewältigt (denn dann hat er sie offenbar nicht nötig), dass aber die Fehlerzahl und mit ihr die Übungsdauer begrenzt bleibt. Das ist umso notwendiger als das Übungsdesign von Orthografietrainer.net nach jedem Fehler Trainingsphasen einlegt. Dadurch steigt die Anzahl zu lösender Aufgaben mit wachsender Fehlerzahl nichtlinear an. Bei einer zu großen Fehlermenge würde dadurch nicht nur die Übung länger als für den optimalen Übungseffekt wünschenswert dauern, sondern auch das Verhältnis von richtig zu falsch gelösten Sätzen gestaltet sich immer ungünstiger. Orthografietrainer.net hat aus den bisher erhobenen Daten und den Rückmeldungen von angemeldeten Fachlehrern ein Limit von sechs Fehlern festgelegt, das möglichst nicht überschritten werden sollte. Unterhalb dieses Limits dauert eine Übung für 68% (Mittelwert + eine Standardabweichung) der Schüler nicht länger als 20 min und das Verhältnis von richtig gelösten zu falsch gelösten Aufgaben liegt höchstens bei 0,12.

Diese Grenze dient Orthografietrainer.net als Richtwert für den Eintrag von Aufgaben, die über die Kompetenztests ermittelt werden. Einem Schüler werden nur solche Aufgaben als Hausaufgaben eingetragen, bei denen statistisch erwartbar ist, dass er unterhalb des Limits von sechs Fehlern bleiben wird. Die statistische Erwartbarkeit ist hierbei dadurch definiert, dass 96% (Mittelwert + zwei Standardabweichungen) der Schüler mit derselben Kompetenz, die diese Aufgabe gelöst haben, unterhalb dieses Limits geblieben sind. Diese Berechnung wird umso aussagekräftiger, je mehr Messwerte vorliegen, je länger und intensiver also Orthografietrainer.net von Lehrern und Schülern genutzt wird.

4.3. Die Schätzung der Personenkompetenz

Während der pair-wise-Algorithmus eine praktikable Möglichkeit bietet, die Schwierigkeit einer Aufgabe zu schätzen, ist die Schätzung der Personenkompetenz in der Item Response

Theorie ein deutlich größeres Problem, da es in der Regel deutlich mehr Testpersonen als Aufgaben gibt und somit die Schwierigkeit jeder Aufgabe lediglich an den anfallenden Datensätzen geschätzt werden kann (nämlich an so vielen, wie Personen am Test teilgenommen haben). Die Personenkompetenz selbst kann hingegen nur anhand der wenigen von der Testperson gelösten Aufgaben ermittelt werden.
Für die Ziele von Orthografietrainer.net ist die Personenkompetenz zwar sekundär (da es ja zunächst um die didaktisch angemessene Anordnung der Aufgaben nach ihrer Schwierigkeit geht), aber dennoch für mehrere Entscheidungsprozesse wichtig: Sie spielt insofern eine Rolle, als es wünschenswert ist, für jeden Trainierenden einen optimalen Startpunkt des Trainings zu ermitteln, von dem aus seine Kompetenzentwicklung erfolgversprechend beginnen kann. Ferner gilt es, anhand der Personenkompetenz abzuschätzen, ob eine Aufgabe, die aufgrund der qualitativen Messungen als sinnvoll für die entsprechende Person eingetragen wurde, von dieser Person auch bewältigt werden kann (vgl. Kap. 4.1). Die Bestimmung der Personenkompetenz eröffnet schließlich die Möglichkeit, den Kompetenzzuwachs nach der Absolvierung einer Übungseinheit zu prüfen. Das ist sowohl für die Einschätzung der Leistung eines Schülers interessant, als auch für die Frage, ob die Zeit reif ist, die nächste Schwierigkeitsstufe anzugehen. Damit ergeben sich für Orthografietrainer.net mittelfristig konkrete Hinweise darauf, wie effizient die verwendeten Übungsmethoden den Kompetenzerwerb unterstützen und inwiefern das Übungsangebot optimiert werden kann.

Der pair-wise-Algorithmus bietet keine Möglichkeit, die Personenfähigkeit zu schätzen und da andere, rechenaufwändigere Bestimmungsmethoden aus technischen Gründen vorerst nicht praktikabel sind, greift Orthografietrainer.net auf didaktische Überlegungen zurück: Sofern das Rasch-Modell für die Rechtschreibkompetenz gilt (zur Diskussion dieser Frage: siehe oben), wird die Lösungswahrscheinlichkeit umso geringer, je schwieriger die Aufgaben sind. Ordnet man also alle vom Schüler gelösten Übungssätze aufsteigend ihrer Schwierigkeit nach, so nimmt die Lösungswahrscheinlichkeit stetig ab, je mehr die Aufgabenschwierigkeit zunimmt. Als Messwert der Personenkompetenz kann dann ein bestimmter, didaktisch begründbarer, Wahrscheinlichkeitswert gelten. Orthografietrainer.net wählt als Grenzwert der Lösungswahrscheinlichkeit den Wert $p=0.9$, und somit eine 90-prozentige Lösungswahrscheinlichkeit, wobei das Verhältnis richtig gelöster Items zur Gesamtzahl der bearbeiteten Aufgaben als Schätzer von p verwendet wird. Im Gegensatz zu anderen Studien ist dieser Wert vergleichsweise hoch angesetzt, hat sich jedoch in Voruntersuchungen als

praktikabel erwiesen. Auch didaktisch gesehen ist es sinnvoll, bei orthografischen Problemen eine hohe Lösungswahrscheinlichkeit als Grenzwert anzusetzen, da ein orthografischer Gegenstand erst dann als beherrscht gelten kann, wenn das entsprechende Rechtschreibthema mit hoher Wahrscheinlichkeit gelöst wird. Da in der konkreten Schreibsituation die Aufmerksamkeitsressourcen von vielen Teilaspekten des Schreibens abhängig sind, muss die Rechtschreibung schon sicher automatisiert sein, um angemessen im freien Schreibprozess angewendet werden zu können. Schulpraktisch lässt sich die 90-Prozent-Grenze dahingehend begründen, da viele gängige Notenschlüssel genau an diesem Punkt die Grenze zwischen der Note „sehr gut“ und „gut“ setzen. Ein Schüler bekommt von Orthografietrainer.net – schulpraktisch betrachtet – die Rechtschreibkompetenz zugewiesen, bei der er in einem Test gerade noch ein „sehr gut“ erreichen würde.

Derzeit ist es kaum möglich, zu entscheiden, ob die Vorgehensweise von Orthografietrainer.net bei der Bestimmung der Personenkompetenz zu adäquaten Messergebnissen führt. Dies wäre dann der Fall, wenn das Rasch-Modell auf die Rechtschreibdaten passt. Passt es nicht, kann es geschehen, dass ein Schüler für schwierige Aufgaben hohe, für leichte hingegen niedrige Lösungswahrscheinlichkeiten aufweist. Dies wäre als Zeichen dafür, dass für diesen Schüler andere Aufgaben leicht bzw. schwer sind als für den Durchschnitt der Nutzer von Orthografietrainer.net und könnte mithin darauf hindeuten, dass eine Vielzahl unterschiedlicher Kompetenzen existiert, deren Ausprägung individuell variiert.

Da sich also die Gültigkeit der Messwerte zur Personenkompetenz derzeit nur auf theoretische Argumente stützt, wird für jeden Schüler zusätzlich der Prozentsatz richtig gelöster Aufgaben angegeben. Sollte das Rasch-Modell Gültigkeit haben, so dürften beide Werte (außer in den Extremscores) in etwa das Gleiche aussagen. Gilt das Rasch-Modell hingegen nicht, so müssten sich häufiger größere Unterschiede zwischen beiden Messwerten auftun. Ein solches Faktum wäre allerdings nicht nur für die Messmethoden von Orthografietrainer.net, sondern auch für die Erforschung des Rechtschreiberwerbs an sich ein wichtiger Befund, würde es doch unter Umständen die gesamte Tradition der Bewertung von Rechtschreibleistungen erheblich infrage stellen.

4.4. Schätzungen bei Über- und Unterforderungen

Die oben vorgestellte Methode zur Ermittlung der Personenkompetenz wird dann problematisch, sobald die 90-Prozent-Grenze der Lösungswahrscheinlichkeit im Test entweder nie erreicht oder nie unterschritten wurde. Im ersten Fall ist der Test für den Schüler

zu schwer, im zweiten ist er zu leicht, um die Kompetenz des Schülers nach der vorgestellten Art und Weise zu ermitteln. Um auch in diesem Falle Kompetenzwerte zur Verfügung stellen zu können, die denen anderer Nutzer vergleichbar sind, muss die tatsächliche Kompetenz anderweitig geschätzt werden. Diese Schätzung kann dann aufgrund der Modellgleichung des Rasch-Modells nicht vorgenommen werden, da sich diese ja nicht nach der Aufgabenschwierigkeit bzw. Personenkompetenz hin auflösen lässt. Als Alternative kommt derzeit nur eine Schätzung aufgrund der Annahme einer linearen Zunahme der Wahrscheinlichkeit infrage, die aber gerade in den Extrembereichen höchst unbefriedigend ist, da sich das Rasch-Modell gerade in den Extremscores vom annähernd linearen Verlauf im Mittelbereich unterscheidet (vgl. Abb. 2). Orthografietrainer.net nähert sich diesem Problem von zwei Seiten: einerseits können die Tests mittelfristig so vervollständigt und interaktiv gestaltet werden, dass das Programm selbst bei Über- bzw. Unterforderungen auf leichtere bzw. schwerere Testsätze ausweicht, andererseits wird mit folgenden Programmversionen auch ein iterativer Rechenalgorithmus zur Verfügung stehen, der die Personenkompetenz nach anderen Methoden ermitteln kann (bspw. nach der UML-Methode, vgl. Rost 2004: 311ff). Bis dahin gilt es, zu überlegen, welche Konsequenzen die unbefriedigenden Schätzungen bei Über- und Unterforderten haben: Aus der wahrscheinlich nicht adäquaten Linearitätsannahme folgt, dass die Kompetenz von Überforderten tendenziell zu gering, die von Überforderten tendenziell zu hoch eingeschätzt wird, sofern das Rasch-Modell gilt. Aus didaktischer Sicht ist dies für die Überforderten keine Gefahr: ihnen werden im schlimmsten Falle Aufgaben noch nicht präsentiert, die sie schon zu lösen imstande wären. Die „Unterforderten“ hingegen könnten tendenziell mit zu schweren Aufgaben konfrontiert werden, als sie zu lösen imstande sind. Da Unterforderung im Test allerdings auch regelmäßig mit dem Eintrag deutlich weniger Aufgaben einhergeht, bildet die Aufgabenmenge hier in gewisser Weise ein Korrektiv. Gleichwohl werden Untersuchungen am Datenmaterial zu klären haben, ob die Behelfswerte für Unterforderte nicht durch andere Berechnungsmethoden abgelöst werden könnten. Bis dahin gilt es, die verantwortlichen Fachlehrer zumindest über die Über- bzw. Unterforderung zu informieren, was Orthografietrainer.net für jeden Test tut.

4.5. Skalierung, Skalenzusammenhänge und Normierung

Orthografietrainer berechnet mehrere unterschiedliche Schwierigkeits- und Kompetenzwerte. Zunächst werden die vier Gebiete der Orthografie (Laut-Buchstaben-Zuordung, Groß- und Kleinschreibung, Getrennt- und Zusammenschreibung, Kommasetzung) als unterschiedliche Kompetenzen angesprochen und getrennt voneinander berechnet. Für jeden Schüler wird

anschließend ein Kompetenzwert in jedem dieser vier Gebiete ermittelt. Darüber hinaus berechnet Orthografietrainer.net eine Gesamtskala, die die einzelnen Aufgaben und Teilgebiete miteinander verrechnet. Anhand dieser Skala wird jedem Schüler ein Kompetenzwert zugewiesen, der seine Rechtschreibkompetenz im Gesamten ausdrückt. Es gilt bei der Interpretation zu beachten, dass der ermittelte Durchschnittswert auf der eigenständigen Berechnung einer Skala nach der pair-wise-Methode beruht – und somit keinesfalls als Durchschnittswert der Einzelskalen betrachten werden kann.

Ersten statistischen Analysen zufolge ähneln sich die beiden Skalentypen (Einzelskalen, Gesamtskala) ihrer Aussage nach sehr deutlich. Die Korrelationen zwischen den Werten je einer Einzelskala und ihrem Pendant in der Gesamtskala liegen zwischen R=.907 (Getrennt- und Zusammenschreibung) und R=.976 (Groß- und Kleinschreibung) bei jeweils höchstem Signifikanzniveau ($p < .000$). Orthografietrainer.net verzichtet dennoch nicht auf die Einzelskalen, weil diese ungleichmäßig auf der Gesamtskala verteilt liegen: Im Mittel sind die Übungssätze zur Laut-Buchstaben-Kombination am leichtesten, gefolgt von der Getrennt- und Zusammenschreibung, der Kommasetzung und schließlich der Groß- und Kleinschreibung. Die Reihenfolge und die Farbgebungen in den Auswertungsdiagrammen von Orthografietrainer.net verdeutlichen diesen Umstand.
Die ermittelten Schwierigkeitswerte wurden auf einen Mittelwert 50 und auf eine Standardabweichung von 20 normiert. Diese Normierung dient lediglich der besseren Lesbarkeit und hat ansonsten keine weiteren Auswirkungen.

5. Fazit: Möglichkeiten und Grenzen der Messmethoden von Orthografietrainer.net

Bei großen statistischen Datenmengen lassen sich auf der Grundlage der Item Response Theorie Trainingspläne erstellen, die qualitativ wie quantitativ auf den Einzelnen abgestimmt sind. Diese Trainingspläne basieren im ersten Schritt auf der qualitativen Aufgabentypisierung anhand der verletzten Rechtschreibregel. In einem zweiten Schritt können auf Grundlage der Einschätzung von Aufgabenschwierigkeiten Voraussagen getroffen werden, welche Person welche Aufgaben innerhalb einer bestimmten Schwierigkeitsstufe bereits zu lösen imstande ist. Somit wird der Trainingserfolg im Voraus einschätzbar.
Mit dieser Vorgehensweise ist es möglich, Trainingspläne zu erstellen, die die konkreten Kompetenzlücken jedes Übenden fokussieren. Anhand der Einschätzung des

Trainingserfolges können bestimmte Rechtschreibthemen bearbeitet oder auch zurückgestellt werden, um die Aufmerksamkeit zunächst anderen Punkten zuzuwenden.

Während der bisherigen Ausführungen ist deutlich geworden dass noch an vielen Stellen theoretische Annahmen und Hypothesen aushelfen müssen, wo klare statistische Fakten wünschenswert wären. Orthografietrainer.net versteht sich als Pilotprojekt, das sich in der Anfangsphase befindet, sodass die derzeit noch störenden Einschränkungen Hinweise für die Weiterentwicklung darstellen können.

Entscheidender Vorteil der Vorgehensweise von Orthografietrainer.net ist, dass das Portal auf Basis der statistischen Daten klare Falsifikationskriterien anbieten kann, wo sich bisherige Erwerbstheorien ausschließlich auf theoretische Plausibilität stützen müssen. Mit zunehmender Nutzung der Plattform und der daraus resultierenden Datenbasis werden Fragen, die heute noch offen sind, beantwortbar und ungeprüfte Hypothesen überprüfbar werden.

Unter den offenen Fragen sind die technischen diejenigen, die an erster Stelle stehen. Die Umsetzung testtheoretischer Methoden im Online-Sektor ist derzeit möglich, erfordert jedoch einen erheblichen Programmier- und Rechenaufwand. Die zeitlichen und finanziellen Ressourcen zu beschaffen, um diese beiden Punkte zu lösen, ist eine der nächsten Aufgaben für die Weiterentwicklung von Orthografietrainer.net.

Mit diesem Entwicklungsschritt werden die aufwändigen Überlegungen zur Ermittlung der Personenkompetenz jeder Testperson wegfallen, die bis dahin als ein Behelf dienen müssen. Ferner werden Tests der Modellgültigkeit möglich, die es erlauben, die bisher getroffenen hypothetischen Annahmen über die orthografische Leistungsentwicklung zu prüfen. Sollten sich die wichtigsten Hypothesen, wie etwa die Gültigkeit des Rasch-Modells für die Daten, als haltbar erweisen, dann entsprechen auch die bis dahin behelfsweise herangezogenen Kompetenzmesswerte den zu erwartenden neuen Messwerten. Sollte die Hypothese der Rasch-Skalierbarkeit hingegen nicht haltbar sein, stehen sowohl die Messwerte von Orthografietrainer.net, als auch nahezu die gesamte Tradition der Rechtschreibübung und Rechtschreibbewertung infrage, da sich die gesamte Rechtschreibpraxis in der Schule implizit auf diese Hypothese stützt.

Selbst bei hinreichender Gültigkeit des Rasch-Modells könnten möglicherweise andere, ggf. mehrdimensionale Modelle die Rechtschreibdaten besser erklären. Auch das hätte bedeutende Auswirkungen auf die Orthografiedidaktik. Zumindest müssten neue Übungen geschaffen oder doch Übungssätze in neuer Weise zu Übungseinheiten zusammengefasst werden. Noch

wichtiger wäre darüber hinaus die Erfassung unterschiedlicher Rechtschreibtypen, die unterschiedliche Anforderungen an das Trainingsmaterial stellen.

Der letzte derzeit noch nicht einschätzbare Punkt betrifft die Feinjustierung in den einzelnen Parametern des Programms, wie etwa das empfohlene Fehlerlimit oder die Grenze, anhand derer im Augenblick die Verrechnung von Personenkompetenz und Aufgabeneintrag geschieht. Orthografietrainer.net stützt sich derzeit auf theoretische Überlegungen, die sich in der Praxis als verwendbar erwiesen haben. Das heißt noch nicht, dass sie bereits das Optimum für den Trainingsprozess darstellen. Vorsichtige Variation der Parameter bei gleichzeitiger kontinuierlicher Messung der statistischen Auswirkungen führt hier auf lange Sicht zu immer besser auf die Bedürfnisse des Einzelnen zugeschnittene Trainingssequenzen, die ggf. in naher Zukunft unterschiedliche Nutzergruppen berücksichtigen können.

Orthografietrainer.net ist bis hierhin nur den ersten Schritt auf einem Weg gegangen, der bedeutsame didaktische Entscheidungen auf ein statistisches Fundament stellt und damit den Anforderungen an eine moderne E-learning-Plattform entspricht.

6. Literatur

AUGST, Gerhard, DEHN, Mechthild: Rechtschreibung und Rechtschreibunterricht. Eine Einführung für Studierende und Lehrende aller Schulformen, 3. Aufl. Stuttgart, Leipzig 2007.

BREDEL Ursula, MÜLLER, Astrid, HINNEY, Gabriele (Hgg.) Schritsystem und Schrifterwerb: linguistisch – didaktisch – empirisch, Belrin, New York 2010.

Dürscheid, Christa: Einführung in die Schriftlinguistik. 2. überarbeitete Aufl., Wiesbaden 2004.

FAY, Johanna, Die Entwicklung der Rechtschreibkompetenz beim Testschreiben. Eine empirische Untersuchung in Klasse 1 bis 4, Frankfurt/M. 2010, zugl Diss. phil. Lüneburg 2009.

GONZALEZ, Eugenio J.: Reporting Student Achievement in Mathematics and Science. In:

GRUND, Martin, HAUG, Gerhard, NAUMANN, Carl Ludwig: DRT 5: Diagnostischer Rechtschreibtest für die 5. Klasse. 2. aktualisierte Aufl. in neuer Rechtschreibung, Göttingen 2004.

GÜNTHER, Hartmut, NÜNKE, Ellen: Warum das Kleine groß geschrieben wird, wie man das lernt und wie man das lehrt. In: Günther, Hartmut, Becker-Mrotzek, Michael: Kölner Beiträge zur Sprachdidaktik 1 2005.

JACHMANN, Michael, Noten oder Berichte? Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern, Opladen 2003.

JÜRGENS, Eiko, Leistung und Beurteilung in der Schule, 6. Auflage St. Augustin 2005.

KUNINA-HABENICHT, Olga, RUPP, André A., WILHELM, Oliver: A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. In: Studies in Educational Evaluation 35 (2009) S. 64–70.

LANGEHEINE, Rolf, PANNEKOEK, Jeroen, VAN DE POL, Frank: Bootstrapping goodness-of-fit measures in categorical data analysis. In: Sociological Methods and Research 24, 4 (1996), S. 492-516).

LINDAUER, Thomas, SCHMELLENTIN, Claudia: Studienbuch Rechtschreibdidaktik. Die wichtigsten Regeln im Unterricht, Zürich 2008.
LÖFFLER, Ilona, MEYER-SCHERPES, Ursula: Probleme beim Erwerb von Rechtschreibkompetenz: Ergebnisse qualitativer Fehleranalysen aus IGLU-E. In: WEINHOLD 2006 a. a. O., S. 199-217.
MAGNO, Carlo: Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. In: The International Journal of Educational and Psychological Assessment Vol. 1,(2009) S. 1-11.
MAND, Johannes: Lese- und Rechtschreibförderung in Kita, Schule und in der Therapie. Entwicklungsmodelle, diagnostische Methoden, Forderkonzepte, Stuttgart 2008.
MARTIN, Michael O., KELLY, Dana L.: TIMSS Technical report Vol. II: Implementation and Analysis, Chestnut Hill 1997, S. 147-174.
MAY, Peter et al.: HSP 1-9. Diagnose orthografischer Kompetenz zur Erfassung der grundlegenden Rechtschreibstrategien, 6. aktualisierte und erweiterte Aufl. Hamburg 2002.
MAYR, Sabine: Grammatikkenntnisse für Rechtschreibregeln? Drei deutsche Rechtschreibwörterbücher kritisch analysiert. Tübingen 2007.
MÜLLER, Hans-Georg: Zum "Komma nach Gefühl". Implizite und explizite Kommakompetenz von Berliner Schülerinnen und Schülern im Vergleich. (=Theorien und Vermittlung der Sprache Bd. 50, hg. von Gerhard Augst et al.), Frankfurt/M. 2007, zugl. Diss. phil. Humboldt-Universität zu Berlin 2007.
PRENZEL, Manfred et al. (Hgg.): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie, Münster 2007.
ROST, Jürgen: Lehrbuch Testtheorie – Testkonstruktion. 2. vollst. überarb. und erw. Aufl. Bern, Göttingen, Toronto, Seattle 2004.
RASCH, Georg: Probabilistic models for some intelligence and attainment tests. Kopenhagen 1960.
RISEL, Heinz: Arbeitsbuch Rechtschreibdidaktik. Baltmannsweiler 2008.
RÖBER-SIEKMEYER, Christa: Ein anderer Weg zur Groß- und Kleinschreibung. Leipzig, Stuttgart, Düsseldorf 1999.
RUPP, André A., TEMPLIN, Jonathan (2009): The (Un)usual Suspects? A Measurement Community in Search of Its Identity. In: Measurement: Interdisciplinary Research & Perspective, 1536-6359, Volume 7, 2, S. 115 – 121.
SACHER, Werner, Leistungen entwickeln, überprüfen und beurteilen, Bewährte und neue Wege für die Primar- und Sekundarstufe, 4. Auflage, Bad Heilbrunn 2004.
SCHEELE, Veronika: Entwicklung fortgeschrittener Rechtschreibfertigkeiten. Ein Beitrag zum Erwerb der „orthographischen“ Strategie, Frankfurt/M. 2006, zugl. Diss. phil. Hannover 2005.
SENATSVERWALTUNG FÜR BILDUNG, WISSENSCHAFT UND FORSCHUNG: Verwaltungsvorschrift Schule Nr. 3/2009, Ergänzung vom 26.08.2009, http://www.berlin.de/imperia/md/content/sen-bildung/rechtsvorschriften/vv_schule_03_2009.pdf?start&ts=1285337189&file=vv_schule_03_2009.pdf (Recherchedatum: 4.10.2010).
THELEN, Tobias: Praktische Möglichkeiten computergestützter Rechtschreibanalyse. In: WEINHOLD 2006, a. a. O., S. 178- 198.
VALTIN, Renate et al.: Orthographische Kompetenzen von Schülerinnen und Schülern der vierten Klasse. In: BOS, Wilfried et al. (Hgg.): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich, Münster 2003, S. 227-264.
WEINERT, Franz E.: Leistungsmessungen in Schulen. Weinheim 2001.
Weinhold, Swantje (Hg.): Schriftspracherwerb empirisch. Konzepte, Diagnose, Entwicklung, Baltmannsweiler 2006.

WILHELM, Oliver, ROBITZSCH, Alexander: Have cognitive diagnostic models delivered their goods? Some substantial and methodological concerns. In: Measurement: Interdisciplinary Research & Perspective, Volume 7, Issue 1 2009, S. 53-57.