Nina Zeuch/Hanneke Geerlings/Heinz Holling/Wim J. van der Linden/ Jonas P. Bertling

Regelgeleitete Konstruktion von statistischen Textaufgaben

Anwendung von linear logistischen Testmodellen und Aufgabencloning

Projekt Regelgeleitete Itementwicklung[1]

1. Einleitung

Neuere Trends in der Bildungsforschung erfordern effektive Kompetenztestung für Diagnose- und Testentwicklungszwecke. Groß angelegte Untersuchungen im Rahmen der Bildungsforschung (wie das Programme for International Student Assessment, PISA oder die Third International Mathematics and Science Study, TIMSS) liefern Ergebnisse und Hinweise für den nationalen und internationalen Vergleich der Lernergebnisse und auch für die Sicherstellung und Verbesserung der Qualität der Lehre. Mit dem steigenden, breit angelegten Gebrauch von bildungsorientierten Tests wächst auch die Notwendigkeit einer effizienteren Testentwicklung und -durchführung. Die Tests sollen möglichst kurz sein und ein Maximum an Informationen über die Kompetenzen der Testperson liefern.

Neuere Entwicklungen im Bereich der Testtechnologie umfassen Versuche der Automatisierung (auf Grundlage theoretischer kognitiver Anforderungen und technischer Formatvorlagen) der Konstruktion von Testaufgaben unter Einbeziehung der Item Response Theorie (IRT) zur Beschreibung der kognitiven Anforderungen der Testaufgaben, der Anwendung computergestützter adaptiver Testung und der Optimierung von Stichproben-Designs. Die Automatisierung der Testaufgabenkonstruktion kann dabei über theoriebasierte Anforderungsprofile und flexible Formatvorlagen erfolgen, die eine Erstellung der Aufgaben durch Computerprogramme ohne Einzelkalibrierung der so konstruierten Aufgaben ermöglichen. Das Grundprinzip des adaptiven Testens ist eine Zuschneidung auf die individuellen Bedürfnisse und Fähigkeiten einer Testperson. Dies erfordert eine fortlaufend aktualisierte Fähigkeitsschätzung während der Testdurchführung sowie große Aufgabenpools oder automatische Aufgabengenerierung (automatic item generation, AIG) und wird im eigentlichen Sinne erst durch computergestütztes Testen ermöglicht (computergestütztes adaptives Testen, CAT, vgl. van der Linden 2003).

Das hier beschriebene Projekt soll diese Entwicklungen aufgreifen, vertiefen und die Ergebnisse integrieren, um ihre Nützlichkeit für die Entwicklung und den Einsatz von

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: HO 1286/5-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

Statistik-Textaufgaben bei OberstufenschülerInnen und UniversitätsstudentInnen zu demonstrieren. Dabei bieten Textaufgaben hervorragende Eigenschaften durch die Verbindung mathematischer Formalismen mit alltäglichen Erfahrungen und die Bedeutung für nahezu jeden Bereich wissenschaftlicher Arbeit sowie für Eingangstests z.B. an Universitäten.

Das übergeordnete Ziel des Projektes ist die Konstruktion eines Softwaresystems, das automatisch eine einzigartige Menge an Aufgaben für alle Proband/innen produzieren und präsentieren kann. Jede folgende Aufgabe sollte an die momentane Fähigkeitsschätzung der Probandin/des Probanden angepasst sein (adaptive Vorgehensweise), weshalb das System in der Lage sein sollte, die Antworten der Probandin/des Probanden zu bewerten und ihre/seine Fähigkeitsschätzung in Echtzeit zu berechnen. Um einzigartige Aufgaben für jede Probandin/jeden Probanden zu generieren, wurden Generierungsregeln für verschiedene Arten von Statistik-Textaufgaben (z.B. Aufgaben zur Wahrscheinlichkeitsrechnung oder zu Konfidenzintervallen) definiert.

In diesem Beitrag wird zunächst ein Überblick über die theoretischen Hintergründe des Projektes in Abschnitt 2 gegeben. Anschließend werden verschiedene IRT-Modelle beschrieben, die zur Bestimmung der Effekte der Generierungsregeln auf die Aufgabenschwierigkeiten herangezogen werden können. Eines dieser Modelle, das linear logistische Testmodell (LLTM), wurde in einer Reihe empirischer Studien eingesetzt, von denen jede zu einer Verfeinerung der aufgabengenerierenden Regeln führte. Beispielhaft werden dazu die Ergebnisse einer Studie zu Konfidenzintervall-Aufgaben in Abschnitt 4.2 beschrieben.

Eine Auswahl der Generierungsregeln für die Wahrscheinlichkeitsaufgaben basiert auf vorherigen Untersuchungen (vgl. Holling/Bertling/Zeuch 2009) und wurde für die Entwicklung eines ersten Prototyps eines automatischen Aufgabengenerators für diesen Aufgabentyp verwendet. Der Artikel schließt mit der Diskussion und einem Ausblick auf künftige Erweiterungen und Anwendungen ab.

2. Theoretischer Hintergrund

Im Folgenden wird die theoretische Grundlage für das vorliegende Projekt dargestellt. Zuerst werden Statistik-Textaufgaben mit wichtigen Merkmalen und Anwendungsbeispielen behandelt. Danach werden automatische Aufgabengenerierung, regelgeleitete Aufgabenkonstruktion und Aufgabencloning erläutert.

2.1 Statistik-Textaufgaben

Statistische Kompetenzen sind ein wichtiger Teil allgemeiner mathematischer Kompetenzen und spielen in nahezu allen wissenschaftlichen Bereichen eine große Rolle. Fundierte Statistikkenntnisse bereiten die Grundlage für wissenschaftliches Arbeiten an der Universität und auch für den Umgang mit Statistik im alltäglichen Leben.

Statistische Kompetenzen können mit diversen Aufgabentypen gemessen werden. Allerdings bieten Textaufgaben mit ihrer Informationsfülle über ein tieferes Verständnis entscheidender statistischer Konzepte über die Beherrschung und den Transfer dieser Kompetenzen über Formeln und Gleichungen hinaus einen idealen Aufgabentyp. Außerdem zeigen Textaufgaben im mathematischen Bereich eine hohe externe Validität, da sie gleichzeitig kreative, logische und mathematische Kompetenzen messen (vgl. z.B. Jonassen 2003). Dies stellt einen wichtigen Anlass dar, sich intensiv wissenschaftlich mit Statistik-Textaufgaben auseinanderzusetzen und damit Statistik einen höheren Stellenwert einzuräumen, ganz im Einklang mit einer der Hauptforderungen der OECD, „Unsicherheit" (eine der vier Subdimensionen mathematischer Kompetenzen bei PISA; vgl. OECD 2003) eine wichtigere Rolle im Bildungsbereich zu verschaffen.

Statistik-Textaufgaben sind eine Unterklasse von mathematischen Textaufgaben. Viele Erkenntnisse über mathematische und vor allem Algebra-Textaufgaben lassen sich auf Statistik-Textaufgaben übertragen, aber letztere sollten als eigenständiger Aufgabentyp betrachtet werden, da verschiedene Unterklassen von mathematischen Textaufgaben qualitativ unterschiedlich sein und deshalb nicht auf einer gemeinsamen konzeptuellen Dimension beschrieben werden können (Arendasy u.a. 2006).

Die Rückführung der Aufgabenschwierigkeit auf bestimmte Konstruktionsregeln und weitere Aufgaben- und Testcharakteristika (Schwierigkeitsmodellierung) für mathematische Textaufgaben wird zunehmend aufwändiger und ambitionierter und umfasst auch Konstruktvalidierung, Aufgabengenerierung und -klassifikation (vgl. Enright/Sheehan 2002). Arendasy u.a. (2006) betrachteten verschiedene Typen von Textaufgaben und entwickelten den Aufgabengenerator Agen, der Vorlagen für die Generierung von Isomorphen (Aufgabenvariationen mit gleicher grundlegender Struktur) nutzt. Dieses Vorgehen ist angelehnt an die Aufgabenproduktion auf der generelleren Basis von Radicals (systematischer Einfluss auf die Aufgabenschwierigkeit) und Incidentals (Oberflächenmerkmale ohne Einfluss auf Aufgabenschwierigkeit). So ist bei einer mathematischen Textaufgabe die zur Aufgabenlösung erforderliche Formel und deren Berechnung als Radical zu betrachten, da sie die Schwierigkeit der Aufgabe beeinflusst; die Rahmengeschichte, in die die Aufgabe eingebettet ist, ist jedoch als Incidental anzusehen, da die Aufgabenschwierigkeit hiervon nicht abhängen dürfte.

Nur sehr wenige Forschungsgruppen beschäftigen sich mit Statistik-Textaufgaben (vgl. z.B. Arendasy u.a. 2006). Regelgeleitete Konstruktion von Statistik-Textaufgaben kann ihren Einsatz in Lehre und Kompetenzmessung unter anderem durch die Möglichkeit von AIG und CAT erleichtern und flexibler gestalten. Hierfür sind vor allem IRT-Modelle wie das LLTM relevant. Derzeit sind uns allerdings keine Ansätze von regelgeleiteter Aufgabenkonstruktion oder Aufgabencloning im Bereich von Statistik-Textaufgaben bekannt.

2.2 Automatische Aufgabengenerierung, regelgeleitete Aufgabenkonstruktion und Aufgabencloning

AIG auf der Basis von theoretisch und empirisch validierten Qualitätskontrollmechanismen dient der Qualitätsverbesserung von Testungen und ermöglicht Aufgabenproduktion unter Minimierung von Fehlerquellen (z.B. uneinheitliche Gestaltung, Tippfehler) und Maximierung der Effizienz, da theoretisch unendliche Aufgabenmengen generiert werden können, sobald das System fertiggestellt ist. Auch wird die Interpretation von Testergebnissen vereinfacht (vgl. Arendasy u.a. 2006). Wenn die bestimmenden Merkmale und Konstruktionsregeln bekannt sind, kann AIG für die Produktion einer großen Anzahl qualitativ hochwertiger Aufgaben genutzt werden. Derzeitige Bemühungen zur AIG lassen sich in die beiden Ansätze der regelgeleiteten Aufgabenkonstruktion und des Aufgabencloning einteilen.

Bei der regelgeleiteten Aufgabenkonstruktion werden die Aufgaben eines Inhaltsbereiches hinsichtlich ihrer kognitiven Anforderungen und schwierigkeitsbestimmenden Merkmale untersucht. Daraus werden Regeln abgeleitet, die diese Strukturen bestimmen (Radicals). Diese Regeln werden in Computeralgorithmen implementiert, welche große Mengen an neuen Aufgaben auf dieser Basis generieren können.

Beim Aufgabencloning wird eine Menge von typischen Aufgaben des Inhaltsbereiches betrachtet, die dann die „Elternaufgaben“ darstellen, aus denen große Familien von „Geschwisteraufgaben“ geklont werden. Normalerweise besteht das Klonen aus der Anwendung von Computeralgorithmen, die unwesentliche Merkmale der Aufgaben (Incidentals) verändern. Die beiden Ansätze sind im Überblick bei Bejar (1993), sowie Irvine und Kyllonen (2002) dargestellt.

3. Statistische Modellierung

Für die Analyse von Daten, die regelgeleitet konstruiert oder durch Aufgabencloning erstellt wurden, wurden verschiedene Modelle entwickelt. In Abschnitt 3.1 werden Modelle beschrieben, die die Generierungsregeln als erklärende Faktoren für die Aufgabenschwierigkeit einbeziehen; in Abschnitt 3.2 wird dargestellt, wie die hierarchische Struktur von Aufgabenpools aus Aufgabencloning in einem Modell berücksichtigt werden kann.

3.1 IRT-Modellierung und das LLTM

Das LLTM gehört zu den IRT-Testmodellen und basiert auf dem Rasch-Modell (RM; vgl. Rasch 1960). Die Grundstruktur dieser Modelle besteht aus einer parametrischen bi- oder multinomialen Verteilung von Antworten von Testpersonen mit Parametern für die Effekte der Testpersonen und der Aufgaben auf die Antwortwahrscheinlichkeiten. Unter anderem können IRT-Modelle auch zur Testung von Hypothesen über mögliche

Problemstrukturen in den Aufgaben und zur Analyse von Antwortdaten von komplexeren Testformen wie adaptive Tests herangezogen werden. Für eine Testperson *j* mit der Fähigkeit θ_j und eine Aufgabe *i* mit der Schwierigkeit σ_j definiert das RM die Wahrscheinlichkeit einer korrekten Antwort ($X_{ij} = 1$) durch:

$$P\left(X_{ij}=1\middle|\theta_j,\sigma_i\right)=\frac{\exp\left(\theta_j-\sigma_i\right)}{1+\exp\left(\theta_j-\sigma_i\right)} \tag{1}$$

Das LLTM zerlegt die Aufgabenschwierigkeit σ_i in $k = 1, \ldots, K$ Basisparameter η_k mit den Gewichten q_{ik} (vgl. Fischer/Molenaar 1995):

$$P\left(X_{ij}=1\middle|\theta_j,q_i,\eta\right)=\frac{\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k\right)}{1+\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k\right)} \tag{2}$$

Die Basisparameter (also die Effekte der Radicals auf die Aufgabenschwierigkeit) werden meistens aufgrund von theoretischen Vorüberlegungen spezifiziert und in einer sogenannten *Q*-Matrix festgehalten, in deren Zeilen die einzelnen Aufgaben und in deren Spalten die Basisparameter stehen. Jeder Aufgabe wird so eine entsprechende Anzahl und Kombination an Basisparametern zugewiesen. Eine Eins in einer Zelle zeigt an, dass in der entsprechenden Aufgabe ein Basisparameter enthalten ist, eine Null, dass der entsprechende Basisparameter nicht enthalten ist. Entweder können vorhandene Aufgaben auf diese Weise klassifiziert werden oder die Aufgaben können nach einer a priori definierten *Q*-Matrix konstruiert werden. Dieses Vorgehen ist auch für die regelgeleitete Aufgabenkonstruktion unerlässlich, bei der Aufgaben entsprechend vorbestimmter kognitiver Strukturen erstellt werden.

Da das LLTM die sehr strikte und oft in der Realität nicht zutreffende Annahme beinhaltet, dass die *Q*-Matrix eine erschöpfende Erklärung für die Aufgabenschwierigkeiten liefert, kann ein zufälliger Fehlerterm in das LLTM einbezogen werden, mit dem Varianzanteile modelliert werden, die nicht durch die spezifizierten Basisparameter erklärt werden (vgl. Janssen/Schepers/Peres 2004). Dieses Modell wird auch Random-Effects LLTM (RE-LLTM) genannt:

$$P\left(X_{ij}=1\middle|\theta_j,q_i,\eta\right)=\frac{\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k+\varepsilon_i\right)}{1+\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k+\varepsilon_i\right)} \tag{3}$$

LLTMs liefern Schätzungen für die vordefinierten Basisparameter, die anzeigen, ob und wenn ja, welche Parameter in welchem Ausmaß einen signifikanten Einfluss auf die Aufgabenschwierigkeiten haben.

Aufgrund der dargestellten Modelleigenschaften eignen sich die LLTMs hervorragend zur Analyse regelgeleitet konstruierter Aufgaben sowie zur Hypothesentestung bezüglich der vermuteten und durch die *Q*-Matrix definierten kognitiven Basisparameter.

3.2 Aufgabencloning-Modelle

Beim herkömmlichen Vorgehen werden die Aufgabenparameter, basierend auf einer Kalibrierung anhand einer Stichprobe von ausreichender Größe, damit der Schätzfehler ignoriert werden kann, auf ihre geschätzten Werte fixiert. Da beim Aufgabencloning die Aufgaben normalerweise in Familien gruppiert sind, die aus der gleichen zugrundeliegenden Struktur abgeleitet sind, scheint ein zufälligkeitsbasierter Ansatz für die Aufgabenparameter eher angemessen. Glas und van der Linden (2003) schlagen ein hierarchisches IRT-Modell für dichotome Aufgaben vor, das diese Gruppierung berücksichtigt. Das Modell behandelt alle Aufgabenparameter im 3-Parameter-Logistischen (3PL)-Modell als zufällig unter Voraussetzung ihrer Familienstruktur. Seien $j = 1, \ldots, J$ die Personen, $f = 1, \ldots, F$ die Aufgabenfamilien, und $i_f = 1, \ldots, I_f$ die Aufgaben der Familie f, so kann das Modell der ersten Ebene als

$$p(X_{i_f j} = 1 \mid \theta_j, a_{i_f}, \sigma_{i_f}, c_{i_f}) = c_{i_f} + (1 - c_{i_f}) \frac{\exp[a_{i_f}(\theta_j - \sigma_{i_f})]}{1 + \exp[a_{i_f}(\theta_j - \sigma_{i_f})]} \tag{4}$$

definiert werden, wobei θ_j, α_{i_f}, σ_{i_f}, c_{i_f} die Fähigkeits-, Diskriminations-, Schwierigkeits-, und Rateparameter sind. Die Aufgabenparameter einer Familie f, als Einheit mit ξ_{i_f} bezeichnet, werden transformiert, sodass ihre Verteilung ausreichend nah an der multivariaten Normalverteilung liegt:

$$\xi_{i_f} \sim MVN(\mu_f, \Sigma_f), \tag{5}$$

mit μ_f und Σ_f als Familienparameter (zweite Ebene).

Als Teil dieses Projektes wurde das Modell erweitert, um eine Erklärung der Aufgabenschwierigkeiten durch die Effekte der angewandten Generierungsregeln zu ermöglichen. Das neue Modell kann somit als Kombination des LLTM mit dem Modell von Glas und van der Linden (2003) angesehen werden. Als Modell der ersten Ebene wurde ein 3-Parameter-Normal-Ogiven (3PNO)-Modell (das bis auf eine Skalierungskonstante annähernd gleich dem 3PL-Modell bei Glas und van der Linden (ebd.) ist) verwendet.

Im neuen Modell wird der Familienschwierigkeitsparameter als eine Kombination aus den Effekten der Radicals η_k angenommen:

$$\mu_{b_f} = \sum_{k=1}^{K} q_{fk} \eta_k \tag{6}$$

Die Variable q_{fk} gibt an, ob Radical k für Familie f benötigt wird. Die Parameter des Modells können in einem Bayesischen Rahmen durch einen Gibbs Sampler geschätzt werden (vgl. Geerlings/van der Linden/Glas eingereicht).

Wenn die Familienparameter mit einem der Aufgabencloning-Modelle geschätzt worden sind, muss prinzipiell eine neu generierte Aufgabe mit bekannter Familienzugehörigkeit nicht mehr kalibriert werden, sondern die bekannten Familienparameter können für die Berechnung der Probandenfähigkeit verwendet werden. Die Genauigkeit der resultierenden Fähigkeitsschätzungen hängt von der Varianz der Aufgabenparameter innerhalb der Familien (Σ_f) ab. Eine erfolgreiche Anwendung des Modells sollte idealerweise in einer großen Varianz der Aufgabenparameter zwischen den und einer geringen Varianz innerhalb der Familien resultieren. Für jeden Probanden/jede Probandin kann dann eine zufällige Aufgabe aus einer Familie (also einer Kombination von Radicals) generiert werden, die optimal bezüglich der aktuellen Fähigkeitsschätzung ist. Gleichzeitig wird der so konstruierte Test durch die Variation der Incidentals aber jedes Mal anders aussehen, was Wiedererkennungseffekte verhindert.

4. Erste Ergebnisse

Zunächst wird die Konstruktion der statistischen Textaufgaben im Rahmen des Projektes dargestellt und es werden erste empirische Ergebnisse beispielhaft aufgezeigt. Daran schließt sich die Darstellung eines Prototyps für einen automatischen Aufgabengenerator an.

4.1 Aufgabentypen und Designprinzipien

Anhand der Lehrpläne für Statistikinhalte im Unterricht der gymnasialen Oberstufe sowie in den Lehrveranstaltungen an Universitäten wurden wichtige basale Operationen der Statistik, die typischerweise Einfluss auf die Lösungswahrscheinlichkeit von Statistik-Textaufgaben haben sollten, identifiziert und als Basisparameter in mehreren Q-Matrizen zur Konstruktion mehrerer Itemmengen definiert. Die Aufgaben wurden halbautomatisch mit Hilfe von LaTex2e-Vorlagen generiert, die durch eine weitestgehend identische Wortwahl und Satzstruktur Missinterpretationen und zusätzliche Fehlervarianz durch unterschiedliches Textverständnis und Satzbaueffekte vermeiden.

4.2 Empirische Studien

Teilmengen der wie oben beschrieben konstruierten Aufgaben wurden in mehreren empirischen Studien mit insgesamt 1274 deutschen OberstufenschülerInnen und PsychologiestudentInnen getestet. Die ersten eingesetzten Aufgaben berücksichtigten vor allem Operationen der grundlegenden Wahrscheinlichkeitstheorie, z.B. den Umgang mit ab-

hängigen oder unabhängigen Wahrscheinlichkeiten. Diese Ergebnisse wurden bereits teilweise veröffentlicht und können im Detail z.B. bei Holling, Bertling und Zeuch (2009) eingesehen werden. Die daraufhin entwickelten Aufgaben widmen sich einem etwas breiteren Inhaltsspektrum (u.a. Varianzanalyse und Konfidenzintervalle). Dabei wurden auch mehrere Testaufgaben zur grundlegenden Wahrscheinlichkeitstheorie nach einem Aufgabencloning-Modell erstellt, diese befinden sich allerdings noch in der Kalibrierungsphase. Beispielhaft soll hier ein Subset von regelgeleitet konstruierten Aufgaben zu Konfidenzintervallen (KI) dargestellt werden, die sich gerade in der Pilotierung befinden. Die einzelnen Aufgaben bestehen aus einer kurzen Rahmengeschichte mit einer daran anschließenden Aufforderung, ein KI oder einen für ein KI benötigten Wert aus einem gegebenen KI zu berechnen.

Der Test besteht aus acht Aufgaben, die unterschiedliche Kombinationen der berücksichtigten Basisparameter VAR (KI für eine Varianz), ANT (KI für einen Anteil), EIN/ZWEI (einseitiges oder zweiseitiges KI), und INV (Inversion der Formel) beinhalten. Wenn die Q-Matrix-Einträge für VAR und ANT Null sind, handelt es sich um ein KI für einen Mittelwert. Abbildung 1 zeigt eine Beispielaufgabe, die die Berechnung eines KIs für einen Anteilswert beinhaltet.

In einer Therapiestudie wird der Frage nachgegangen, ob ein neuartiges Therapieprogramm effektiv im Sinne der Befindlichkeitsverbesserung der Patienten ist. 72 der insgesamt 120 Teilnehmer berichten eine deutliche Besserung. Die Klinik verspricht aber, dass mit diesem Programm mindestens 50 Prozent der Patienten eine Verbesserung der Befindlichkeit erfahren.
Berechnen Sie ein Konfidenzintervall für die Ergebnisse der Studie, das die Klinikleitung zur Überprüfung ihres Versprechens heranziehen könnte, wenn eine Sicherheit von 90 Prozent berücksichtigt werden soll.

Abb. 1: Beispielaufgabe Konfidenzintervall-Test

Aufgabe	VAR	ANT	EIN/ZWEI	INV
1	0	0	0	1
2	0	1	0	0
3	1	0	0	1
4	0	0	1	0
5	0	1	1	0
6	1	0	1	0
7	0	0	1	1
8	0	1	1	1

Anmerkungen: VAR = „Varianz“, ANT = „Anteilswert“, EIN/ZWEI = „ein- oder zweiseitig“, INV = „Inversion“.

Tab. 1: Q-Matrix für den Konfidenzintervall-Test

Tabelle 1 zeigt die *Q*-Matrix für die acht Aufgaben (die Beispielaufgabe aus Abbildung 1 entspricht Aufgabe 2 i nder Designmatrix).

Die Aufgaben wurden 86 PsychologiestudentInnen der ersten beiden Fachsemester an der Westfälischen Wilhelms-Universität Münster vorgelegt. Durchschnittlich wurden 3,59 (45 Prozent) der acht Aufgaben korrekt beantwortet. Die Aufgabenschwierigkeiten bewegen sich zwischen 0,28 für Aufgabe 8 und 0,69 für Aufgabe 4. Die interne Konsistenz ist mit einem Cronbachs Alpha von 0,48 sehr gering. Der *Q*-Index weist mit Werten zwischen 0,14 für Aufgabe 7 und 0,21 für Aufgabe 5 einen guten Rasch-Modellfit für alle Aufgaben auf (vgl. Rost/von Davier 1994). Tabelle 2 zeigt die sehr ähnlich ausfallenden LLTM- und RE-LLTM-Schätzungen.

		LLTM		**RE-LLTM**	
	Parameter	**Schätzung**	**SE**	**Schätzung**	**SE**
	Konstante	0.36	0.25	0.37	0.35
Feste Effekte	VAR	−0.20	0.22	−0.20	0.31
	ANT	−0.63**	0.20	−0.64*	0.29
	EIN/ZWEI	0.11	0.18	0.11	0.26
	INV	−0.76**	0.18	−0.77**	0.26
Zufällige Effekte	θ_j	0.51	0.19	0.49	0.19
	ε_i	–	–	0.06	0.06

Anmerkungen: $*p < .05$. $**p < .01$. SE = Standardfehler. VAR = „Varianz“, ANT = „Anteilswert“, EIN/ZWEI = „ein- oder zweiseitig“, INV = „Inversion“.

Tab. 2: Parameterschätzungen für LLTM und RE-LLTM im Konfidenzintervall-Test

Zwei der vier Basisparameter (ANT und INV) sind sowohl im LLTM als auch im RE-LLTM statistisch signifikant und haben somit einen inkrementellen Einfluss auf die globale Aufgabenschwierigkeit. Ein Likelihood-Ratio-Test konnte keinen Vorteil des RE-LLTM gegenüber dem LLTM nachweisen.

Die Korrelation zwischen LLTM- und Rasch-Aufgabenparametern beträgt 0,79. Daraus ergibt sich eine gute Varianzaufklärung von $R^2 = 0{,}63$.

Diese Resultate des KI-Tests sind, wahrscheinlich aufgrund der geringen Testlänge und Stichprobengröße, nicht einwandfrei. Die Tendenz ist dennoch vielversprechend, konnten doch zwei statistisch signifikante Basisparameter identifiziert werden. Diese Aufgabenform soll weiterentwickelt und an größeren Stichproben getestet werden.

4.3 Automatischer Aufgabengenerator

Ein automatischer Aufgabengenerator wird für Statistik-Textaufgaben entwickelt, in dem Operationen der grundlegenden Wahrscheinlichkeitstheorie geprüft werden. Die Aufgaben ähneln denen, die in Holling, Bertling und Zeuch (2009) dargestellt werden.

Jede Statistik-Textaufgabe besteht aus einer Rahmengeschichte, die die relevanten numerischen Informationen für die Antwortberechnung enthält und einer Frage, die die Berechnung einer bedingten Wahrscheinlichkeit („ua“), eines Komplementärereignisses („nicht“), einer Wahrscheinlichkeit für eine Schnittmenge („uu“) oder einer Wahrscheinlichkeit für eine Verbundmenge („oder“) erfordert. Die Struktur der Kontextgeschichten ist für jede Frage gleich. Die einzige Kontextvariation wird durch die Incidentals verursacht, die Informationen über Subjekt und Objekt der Geschichte und die Interpretation der verwendeten Variablen liefern. Eine Frage, die beispielsweise die Berechnung einer Gegen-, bedingten und Verbundmengen-Wahrscheinlichkeit erfordert, kann durch die Anwendung einiger weniger Aussagen (siehe Abbildung 2) generiert werden. Eine Sammlung von Subjekten, Objekten und Variablen wird zur Erzeugung der Oberflächenunterschiede zwischen den Aufgaben verwendet.

“Wie groß ist die Wahrscheinlichkeit, dass”	Wie groß ist die Wahrscheinlichkeit, dass
<*Subjekt-Artikel*> <*Subjekt*> „ein/eine/einen“	der Buchhändler ein
<*Objekt Singular*> „hat“, <*Relativpronomen*>	Buch hat, das
if(not=1) “nicht”	nicht
if(uu=1 & oder!=1) “sowohl”	
if(oder=1) „entweder“	entweder
„ein/eine/ein“ <*Merkmalsausprägung 1*> <*Variable 1*> <*Verb 1*>	einen gelben Umschlag hat
if(oder=1) „oder“	oder
if(oder=1 & uu=1) „sowohl“	
„ein/eine/ein“ <*Merkmalsausprägung 1*> <*Variable 1*> <*Verb 1*>	einen grünen Umschlag hat,
if(ua=1) „ , vorausgesetzt, ‘‘ <*Artikel*>	vorausgesetzt, das
<*Objekt Singular*> <*Verb 3*> „ein/eine“einen“	Buch hat eine
<*Merkmalsausprägung 3*> <*Variable 3*>	männliche Hauptperson
if(uu=1) „als auch ein/eine/ein“	
<*Merkmalsausprägung 2*> <*Variable 2*>	
<*Verb 2*>	

Abb. 2: Vereinfachte Struktur der Fragen und eine Beispielfrage

5. Diskussion und Ausblick

Im vorliegenden Projekt werden Inhalte der kognitiven Psychologie, Psychometrie und Computerwissenschaften kombiniert, um ein Testsystem für Statistik-Textaufgaben zu entwickeln, das einen adaptiven, einzigartigen Test für jeden Probanden/jede Probandin erschaffen kann.

Die ersten Projektergebnisse sind durchweg vielversprechend. So konnten verschiedenste kognitive Komponenten identifiziert und in halbautomatischer Aufgabenkonstruktion als Vorstufe zur vollautomatischen Generierung mit Hilfe von Textbausteinen umgesetzt werden. Beispielhaft wurde ein Test zu Konfidenzintervallen dargestellt. Diese Ergebnisse und die aller weiteren empirischen Untersuchungen zeigen die Anwendbarkeit von regelgeleiteter und automatischer Aufgabengenerierung auf Textaufgaben mit Statistikinhalten. Die regelgeleitete Konstruktion wurde in mehreren Aufgabenmengen umgesetzt und empirisch überprüft. Die Ergebnisse dienen nun zur Verfeinerung des automatischen Aufgabengenerators. Es konnten in LLTM-Analysen verschiedene signifikante schwierigkeitsgenerierende Merkmale identifiziert werden. Außerdem zeigen die konstruierten Aufgaben einen guten RM-Fit und stellen den Ausgangspunkt für verfeinerte inhaltliche und konstruktionstechnische Weiterentwicklungen und adaptive Implementierungen dar. Verschiedene mögliche Modelle für die Kalibrierung der automatisch generierten Aufgaben wurden aufgezeigt. Eine wichtige Frage ist, ob bessere Modellpassung die höhere Parameteranzahl im komplexeren hierarchischen IRT-Modell rechtfertigt. Es ist eine Studie geplant, in der die Passung aller hier erwähnten Modelle verglichen wird: Das LLTM, das Aufgabencloning-Modell von Glas und van der Linden (2003) sowie die erweiterte Version dieses Modells (beschrieben in Abschnitt 3.2). Die Modelle sollen mit Hilfe der gewonnenen Daten aus den empirischen Erhebungen zu Statistik-Textaufgaben aus 4.3 verglichen werden.

Literatur

Arendasy, M./Sommer, M./Gittler, G./Hergovich, A. (2006): Automatic generation of quantitative reasoning items: A pilot study. In: Journal of Individual Differences 27, S. 2–14.

Bejar, I. (1993): A generative approach to psychological and educational measurement. In: Frederiksen, N./Mislevy, R.J./Bejar, I.I. (Hrsg.): Testtheory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates, S. 323–357.

Enright, M.K./Sheehan, K.M. (2002): Modeling the difficulty of quantitative reasoning items: Implications for item generation. In: Irvine, S.H./Kyllonen, P.C. (Hrsg.): Item generation for test development. Mahwah, NJ: Erlbaum, S. 129–157.

Fischer, G.H./Molenaar, I.W. (Hrsg.) (1995): Rasch Models. Foundations, Recent Developments, and Applications. New York: Springer.

Geerlings, H./van der Linden, W.J./Glas, C.A.W. (eingereicht): Modeling rule-based item generation.

Glas, C.A.W./van der Linden, W.J. (2003): Computerized adaptive testing with item cloning. In: Applied Psychological Measurement 27, S. 247–261.

Holling, H./Bertling, J.P./Zeuch, N. (2009): Probability word problems: Automatic item generation and LLTM modelling. In: Studies in Educational Evaluation 35, S. 71–76.

Irvine, S.H./Kyllonen, P.C. (2002): Item generation for test development. Mahwah, NJ: Erlbaum.

Janssen, R./Schepers, J./Peres, D. (2004): Models with item and item group predictors. In: De Boeck, P./Wilson, M. (Hrsg.): Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. New York: Springer, S .189–210.

Jonassen, D.H. (2003): Designing research-based instruction for story problems. In: Educational Psychology Review 15, S. 267–296.

OECD (Hrsg.) (2003): The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills. Paris: OECD.
Rasch, G. (1960): Probabilistic Models for some Intelligence and Attainment tests. Copenhagen: Pædagogiske Institut.
Rost, J./von Davier, M. (1994): A conditional item fit index for Rasch models. In: Applied Psychological Measurement 18, S. 171–182.
Van der Linden, W. (2003): Some new developments in adaptive testing technology. In: Journal of Psychology 216, S. 3–11.

Anschrift der Autor/innen

Dipl.-Psych. Nina Zeuch, Westfälische Wilhelms-Universität Münster,
Lehrstuhl für Statistik und Methoden, Fliednerstraße 21, D-48149 Münster
E-Mail: n_hoff01@uni-muenster.de

Prof. Dr. Heinz Holling, Westfälische Wilhelms-Universität Münster,
Lehrstuhl für Statistik und Methoden, Fliednerstraße 21, D-48149 Münster
E-Mail: holling@uni-muenster.de

Dipl.-Psych. Jonas P. Bertling, Westfälische Wilhelms-Universität Münster,
Lehrstuhl für Statistik und Methoden, Fliednerstraße 21, D-48149 Münster
E-Mail: jonas.bertling@uni-muenster.de

MSSc. Hanneke Geerlings, University of Twente, Department of Research Methodology,
Measurement and Data Analysis, Faculty of Behavioral Sciences, P.O. Box 217,
7500 AE Enschede, The Netherlands
E-Mail: h.geerlings@gw.utwente.nl

Prof. Dr. Wim J. van der Linden, University of Twente, Department of Research Methodology,
Measurement and Data Analysis, Faculty of Behavioral Sciences, P.O. Box 217,
7500 AE Enschede, The Netherlands
E-Mail: w.j.vanderlinden@utwente.nl