Kompetenzzentrum für Bildungsevaluation und Leistungsmessung an der Universität Zürich · KBL
Centre de compétences en évaluation des formations et des acquis à l'Université de Zurich · CEA
Competence Centre for Educational Evaluation and Assessment at the University of Zurich · CEA

Stellwerk: ein computergestütztes adaptives Testsystem

Testtheoretische Grundlagen und erste Erfahrungen

Urs Moser
Zürich, Dezember 2006

INHALT

1 Einleitung 3

1 Testtheoretische Grundlagen 3

1.1 Stellwerk: ein wissenschaftliches Testverfahren für die Schule 3

1.2 Stellwerk: ein computergestütztes adaptives Testsystem 4

1.3 Testtheoretische Grundlagen 6

2 Kalibrierung der Testaufgaben und Normierung 10

2.1 Testdesign 10

2.2 Zusammenstellung der Tests 11

2.3 Stellwerkskala 12

3 Ergebnisse der Normierung 2005 14

3.1 Mathematik 14

3.2 Deutsch 16

3.3 Französisch 18

3.4 Englisch 19

3.5 Natur und Technik 20

4 Funktionsweise des adaptiven Testens im Jahr 2006 23

4.1 Testergebnisse im Jahr 2006 23

4.2 Häufigkeit der Aufgabenzuteilung durch den Algorithmus und Schwierigkeitsparameter der Aufgaben 24

5 Ausblick 25

6 Literatur 25

1 Einleitung

Im Auftrag des Lehrmittelverlags des Kantons St. Gallen wurden für Stellwerk folgende Aufgaben ausgeführt:

- Konzeption des Testdesigns unter Berücksichtigung der notwendigen Verbindung unterschiedlicher Testversionen durch Link-Aufgaben
- Zufällige Zuteilung der Testaufgaben zu den Schülerinnen und Schülern für die Kalibrierung der Testaufgaben beziehungsweise für die Normierung der Tests
- Skalierung der Daten auf der Grundlage der Item-Response-Theorie beziehungsweise der probabilistischen Testtheorie (Rasch-Modell)
- Schätzung der Itemparameter nach dem Rasch-Modell zur Steuerung der Testaufgaben mit Hilfe eines Algorithmus
- Schätzung der Personenparameter nach dem Rasch-Modell zur Darstellung der Ergebnisse der Schülerinnen und Schüler der Normierung im Jahr 2005
- Entwicklung und Programmierung eines Algorithmus für adaptives Testen auf der Grundlage des Rasch-Modells und für die Schätzung der Personenparameter
- Wissenschaftliche Beratung bei der Entwicklung des Testsystems «Stellwerk»

Der vorliegende Bericht informiert über die testtheoretischen Grundlagen der Stellwerk-Tests, über die Ergebnisse der Normierung der Tests im Jahr 2005 sowie über erste Erfahrungen mit dem computergestützten adaptiven Testsystem im Jahr 2006.

1 Testtheoretische Grundlagen

1.1 Stellwerk: ein wissenschaftliches Testverfahren für die Schule

Mit Stellwerk stellt der Lehrmittelverlag des Kantons St. Gallen der Schule Tests zur Verfügung, mit denen die Schülerinnen und Schüler ihre Leistungen in verschiedenen Kompetenzbereichen am Computer ausweisen können. Weshalb braucht es dazu Tests? Was ist das Besondere an Tests und worin unterscheiden sich wissenschaftlich konzipierte Tests im Vergleich zu gewöhnlichen Prüfungen oder informellen Tests, die Lehrpersonen im Schulalltag einsetzen? Die folgende Definition von Tests stammt von Gustav Lienert, dem Verfasser eines deutschsprachigen Grundlagenwerks zum Thema «Testaufbau und Testanalyse», und umfasst die wichtigsten Merkmale eines Tests.

> «*Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung*» (Lienert, 1969, S.7).

Die Definition von Gustav Lienert ist auch heute noch brauchbar, weshalb sie in aktuellen Lehrbüchern über Testtheorie und Testkonstruktion immer noch diskutiert wird (Rost, 2003, S. 17). Ein Test ist ein *Routineverfahren*, das heisst, das Verfahren ist in Bezug auf die Durchführbarkeit und Auswertung bereits an einer Stichprobe erprobt beziehungsweise normiert worden. Das Routineverfahren wird als *wissenschaftlich* bezeichnet, weil es auf einer Testtheorie beruht, die Annahmen darüber macht, unter welchen Bedingungen aus den Testergebnissen welche Aussagen über die getesteten Personen abgeleitet werden können. Der Test bezieht sich auf ein *Persönlichkeitsmerkmal*, also auf die Erfassung eines relativ stabilen und konsistenten Merkmals einer Person, das für die Testleistung verantwortlich ist. Dieser Teil der Definition ist allerdings etwas zu eng gefasst, weil in der Regel zwischen Leistungs- und Fähigkeitstests einerseits und Persönlichkeitstests und Einstellungsfragebogen andererseits unterschieden wird (Leutner, 2001, S. 525). Tests haben zudem zum Ziel, möglichst *quantitative Aussagen* über die Ausprägung des Persönlichkeitsmerkmals beziehungsweise der Fähigkeit zu machen (Rost, 2003). Auch hier scheint die Definition etwas zu eng zu sein, weil die Aussagen über ein Testergebnis auch qualitativ sein können, indem konkret umschrieben wird, über welche Fähigkeiten eine Person verfügt.

Die drei Elemente der Definition von Lienert (1969) – wissenschaftliches Routineverfahren, empirisch abgrenzbares Persönlichkeitsmerkmal und quantitative Aussage über die Merkmalsausprägung – waren auch für die Entwicklung der Stellwerk-Tests relevant. Stellwerk-Tests sind Routineverfahren, die für die 8. Klasse entwickelt und im Jahr 2005 bei über 6000 Schülerinnen und Schülern normiert wurden. Stellwerk bietet zurzeit Tests für die Kompetenzbereiche Mathematik, Naturwissenschaften (Biologie, Chemie und Physik) Sprachen (Deutsch, Englisch und Französisch) sowie Vorstellungsvermögen an. Das Testsystem kann für unterschiedlichste Kompetenzbereiche auf verschiedenen Bildungsstufen genutzt werden, sofern dazu geeignete Testaufgaben unter Berücksichtigung der testtheoretischen Voraussetzungen entwickelt und die Tests normiert werden. Stellwerk führt zu quantitativen, aber vor allem auch zu qualitativen Aussagen über die Fähigkeiten der Schülerinnen und Schüler. Das heisst, die Testergebnisse können nicht nur im sozialen Vergleich interpretiert werden, sondern auch im Vergleich zu Deskriptoren, Begriffen und Beispielaufgaben, wie sie im Referenzrahmen und in den Interpretationshilfen zu Stellwerk ausgewiesen sind (www.stellwerk-check.ch).

1.2 Stellwerk: ein computergestütztes adaptives Testsystem

Stellwerk ist der Name für ein Testsystem, das Tests online über den Computer anbietet, die adaptiv über einen Algorithmus gesteuert werden. Computergestütztes Testen bedeutet, dass die Testaufgaben am Bildschirm erscheinen und von einer Person mit Hilfe der Tastatur oder der Maus gelöst werden. Die Korrektur der Testaufgaben entfällt beziehungsweise wird vom Computer ausgeführt. Diese unmittelbare Ergebnisrückmeldung wird der Person zwar nicht mitgeteilt, jedoch dafür genutzt, den Test zu steuern. Die Stellwerk-Tests sind dynamische, computergestützte adaptive Tests, die sich den Fähigkeiten jener Person anpassen, die den Test bearbeitet.

Bei einem konventionellen Test, meist ein so genanntes Papier-Bleistift-Verfahren – beispielsweise die Tests, wie sie Klassencockpit anbietet –, werden vielfach allen Personen die gleichen Aufgaben vorgelegt, oder die Tests enthalten zwar unterschiedliche

Aufgaben, sind aber alle in etwa gleich schwierig. Schwierige Aufgaben einer Person mit geringen Fähigkeiten vorzulegen, kann als Zeitverschwendung bezeichnet werden, die meist noch mit einer Frustration der Person verbunden ist. Umgekehrt langweilen sich Personen mit sehr grossen Fähigkeiten, wenn sie einfache Aufgaben bearbeiten müssen.

Damit sich der Test den Fähigkeiten einer Person anpassen kann, wird die Auswahl der Testaufgaben durch einen Algorithmus so gesteuert, dass jeweils die Schwierigkeit einer Aufgabe möglichst nahe bei der gemessenen Fähigkeit liegt, die fortwährend aufgrund des Lösungsverhaltens der Person neu berechnet wird. Sobald sich bei der Berechnung der Fähigkeit der Person nahezu keine Änderungen mehr einstellen, wird der Test beendet. Weil bei diesem Prozess die Lösungen vom Computer umgehend als richtig oder falsch codiert werden müssen, eignen sich für adaptive Tests Aufgaben im Multiple-Choice-Format (Mehrfachwahlantwort) besonders gut. Es werden aber auch offene Aufgaben eingesetzt, deren Lösungen eine kurze schriftliche Antwort verlangen und vom Computer einwandfrei korrigiert werden können.

Ein Test beginnt für alle Personen mit einer relativ einfachen, zufällig ausgewählten Aufgabe. Nachdem die Aufgabe gelöst wurde, schätzt das System aus dem Schwierigkeitsparameter der Aufgabe und der Lösung (richtig oder falsch) die Fähigkeit der Person (Personenparameter). Danach sucht das System jene Aufgabe, deren Schwierigkeitsparameter am nächsten bei der geschätzten Fähigkeit der Person beziehungsweise dem Personenparameter liegt. Löst beispielsweise eine Person alle Aufgaben von Beginn an richtig, dann schlägt sich dies in der Schätzung ihrer Fähigkeit nieder. Der Personenparameter wird grösser, und dementsprechend weist das System der Person schwierigere Aufgaben zu. Umgekehrt sinkt der Personenparameter, wenn die Person die Aufgaben falsch löst. Das System weist der Person in diesem Fall Aufgaben zu, deren Schwierigkeitsparameter kleiner sind. Die Schätzung der Fähigkeit weicht von der wahren Fähigkeit, die sich ja nicht beobachten, sondern nur aus dem Testergebnis erschliessen lässt, immer weniger ab. Der Test dauert so lange, bis grössere Schwankungen bei der Schätzung der Fähigkeit ausbleiben und das System nur noch Aufgaben zuweist, deren Schwierigkeitsparameter sich kaum mehr von den geschätzten Personenparametern unterscheiden. Die letzte Schätzung des Personenparameters entspricht dem Gesamtwert im Test.

Adaptive Tests haben den Vorteil, dass sie sich relativ rasch den Fähigkeiten der Schülerinnen und Schüler anpassen. Dabei ist es wünschenswert, dass der Aufgabenpool möglichst gross und in der Schwierigkeit möglichst breit streut. Schwache Schülerinnen und Schüler werden dann nicht mit zu schwierigen Aufgaben frustriert, starke werden nicht mit zu einfachen Aufgaben gelangweilt, denn sowohl zu schwierige als auch zu einfache Aufgaben liefern im Grunde genommen keine Informationen zur Schätzung der Kompetenz. Adaptive Tests sind aus diesem Grunde ökonomisch und führen zugleich zu einer sehr zuverlässigen Schätzung der Kompetenzen, weil vor allem jene Aufgaben bearbeitet werden, die für das Fähigkeitsniveau der einzelnen Personen eine hohe Messgenauigkeit (Aufgabeninformation) aufweisen (Amelang & Schmidt-Atzert, 2006, S. 81, Kubinger, 2003, S. 6.).

1.3 Testtheoretische Grundlagen

Testtheorien

Um zu einer wissenschaftlichen Aussage über die Ausprägung eines Persönlichkeitsmerkmals zu gelangen, basieren Tests auf einer Testtheorie, die den Zusammenhang zwischen dem Merkmal, beispielsweise mathematische Kompetenz, und dem Testergebnis, beispielsweise Anzahl Punkte, beschreibt. Eine Testtheorie ist ein statistisches Modell, das beschreibt, wie von den Testergebnissen auf das Persönlichkeitsmerkmal oder die Fähigkeiten der Person geschlossen werden kann. Die Testtheorie sagt also primär nichts über die verschiedenen Arten von Tests oder über Konstruktionsprinzipien aus, sondern beschäftigt sich insbesondere mit dem Zusammenhang von Testverhalten oder Testergebnis und dem gemessenen Merkmal (Persönlichkeitsmerkmal oder Fähigkeit). Grundsätzlich lassen sich zwei verschiedene Theorien unterscheiden: die klassische Testtheorie und die Item-Response-Theorie.

Klassische Testtheorie

Die überwiegende Zahl von Tests sind nach den Regeln der klassischen Testtheorie konzipiert worden. Ein typisches Beispiel sind die Tests, die im Rahmen von Klassencockpit entwickelt werden. Bei der klassischen Testtheorie handelt es sich um eine Messfehlertheorie, die auf die erhaltenen Messwerte (Testergebnisse) angewendet wird, um deren Fehleranteil zu bestimmen. Der Messfehler umfasst die Gesamtheit aller unsystematischen und nicht kontrollierbaren oder vorhersagbaren Einflussgrössen, die auf das Testergebnis einwirken können, beispielsweise Müdigkeit oder Prüfungsangst (Amelang & Schmidt-Atzert, 2006). Das Testergebnis setzt sich folglich aus dem wahren Wert, beispielsweise der gemessenen Mathematikkompetenz, und dem Messfehler, beispielsweise der Tageszeit, zusammen.

Damit der Messfehler möglichst klein gehalten werden kann, muss eine objektive Durchführung des Tests gewährleistet sein. Das heisst, dass das Testergebnis unabhängig von der durchführenden und auswertenden Person zustande kommt. Die Standardisierung der Durchführung und Auswertung ist bei einem computergestützten Test besonders hoch, weil beides elektronisch verläuft. Allerdings setzen computergestützte Testverfahren einen einwandfreien Betrieb des Systems voraus. Zudem darf eine Person bei den Testaufgaben nicht aufgrund mangelnder Computerkenntnisse scheitern.

Sofern das Testergebnis zudem ein zuverlässiger Indikator für die Fähigkeit einer Person ist, muss der Test genau messen beziehungsweise reliabel sein. Das heisst, dass eine Testwiederholung unter gleichen Bedingungen, also ohne Einflüsse durch Übung und Gedächtnis, zu einem gleichen Ergebnis führen muss[1]. Ein Test gilt als reliabel, wenn er das, was er zu messen vorgibt, möglichst messfehlerfrei misst (Leutner, 2001, S. 527).

Zur Bestimmung der Messgenauigkeit (Reliabilitätskoeffizient) gibt es verschiedene Methoden, beispielsweise kann der Test wiederholt werden (Testwiederholung), in zwei Versionen vorgelegt werden (Paralleltests) oder in zwei Teile zerlegt werden (Testhalbie-

[1] Die Annahme, dass eine Testwiederholung zu gleichen Ergebnissen führen muss, ist nur bei sehr kurzen Zeiträumen vertretbar und für Leistungstests kaum haltbar (Amelang & Schmidt-Atzert, 2006, S. 60).

rung). In jedem Fall wird überprüft, wie gut die Testergebnisse übereinstimmen. Durch die Verallgemeinerung der Halbierungsmethode wird der Test nicht nur in zwei Hälften zerlegt, sondern in so viele Teile, wie der Test Aufgaben enthält. Entsprechend werden die Übereinstimmungen ermittelt. Das Ziel der Testentwicklung ist es, den Anteil des Messfehlers zu bestimmen und durch optimale Aufgabenselektion zu reduzieren.

Mit Hilfe des Reliabilitätskoeffizienten kann der Standardmessfehler berechnet werden. Der Standardfehler ist jener Anteil des Testergebnisses, das auf die unvollständige Perfektion bei der Messung zurückzuführen ist. Ist der Standardmessfehler bekannt, dann kann festgestellt werden, in welchem Vertrauensbereich der wahre Wert liegt. Für die Interpretation eines Testergebnisses ist eminent wichtig, dass die Grundannahmen der klassischen Testtheorie berücksichtig werden: Bei einem Testergebnis handelt es sich nicht einfach um einen wahren Wert, sondern um ein fehlerbehaftetes Messergebnis. Das Testergebnis bewegt sich in einem Vertrauensbereich, innerhalb jenem der wahre Wert einer Person vermutet wird.

Der Vorteil des computergestützten adaptiven Testens liegt darin, dass der Vertrauensbereich des wahren Testwertes gering ist. Der Test wird erst dann abgebrochen, wenn der Vertrauensbereich des wahren Testergebnisses eine bestimmte Schwelle unterschritten hat. Die Genauigkeit der Messung wird auch deshalb erhöht, weil sich der Test den Fähigkeiten anpasst, was zur Folge hat, dass kaum Aufgaben gelöst werden, die viel zu schwierig oder viel zu einfach sind. Solche Aufgaben führen zu keinen brauchbaren Informationen.

Testtheoretische Bedingungen für adaptives Testen

Eine der bedeutsamsten Unzulänglichkeiten der klassischen Testtheorie besteht darin, dass die geschätzten Parameter populations- oder stichprobenabhängig sind. Beispielsweise wird die Schwierigkeit einer Aufgabe definiert als die Anzahl richtiger Lösungen einer Aufgabe in einer Stichprobe. Das heisst, dass die berechneten Testwerte der Personen, aber auch die Aufgaben- und Teststatistiken (beispielsweise Schwierigkeit einer Aufgabe, Reliabilitätskoeffizient, Standardmessfehler) jeweils von der Personenstichprobe beziehungsweise von der Referenzpopulation abhängen. Je nach Stichprobe ist eine Aufgabe folglich schwierig oder einfach. Die Interpretation der Testergebnisse kann deshalb nur über den Vergleich von Testergebnissen untereinander, also normbezogen, erfolgen. Damit dieser Vergleich sinnvoll möglich wird, muss jede Person die gleiche und vollständige Auswahl an Testaufgaben bearbeitet haben. Genau dies soll aber beim adaptiven Testen verhindert werden. Jeder Person soll eine ihren Fähigkeiten entsprechende Auswahl von Aufgaben vorgelegt werden.

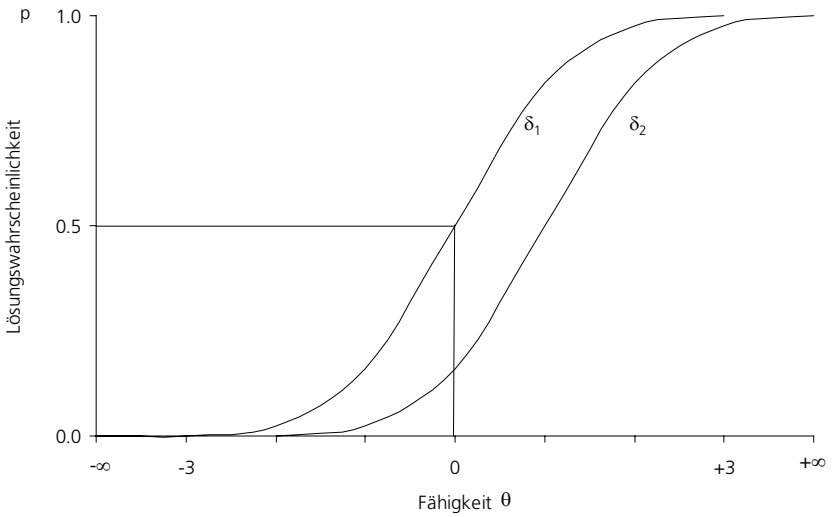
Zudem wird bei der klassischen Testtheorie davon ausgegangen, dass der Messfehler für alle Personen konstant ist, obwohl er bei Personen im mittleren Merkmalsbereich am geringsten, für solche in den beiden extremen Merkmalsbereichen am grössten ausfällt. Weil der adaptive Test Aufgaben auswählt, die sozusagen für jede Person im mittleren Merkmalsbereich liegen, können die Schätzungen genauer erfolgen als wenn Tests eingesetzt werden, die jeweils von einer gesamten Population gelöst werden. Tests, die auf der Basis der klassischen Testtheorie entwickelt werden, sind deshalb für manche Personen entweder zu lang oder aber zu ungenau. Zudem fallen je nach Homogenität oder Heterogenität der Personenstichprobe die Reliabilitätskoeffizienten hoch oder niedrig aus. Die Reliabilität ist nicht für alle Teilstichproben gleich gross.

Die Item-Response-Theorie

Tests auf der Basis der Item-Response-Theorie haben in den letzten Jahren durch internationale Schulleistungsmessungen – und in der Schweiz speziell durch die damit verbundene Absicht, Bildungsstandards einzuführen – Auftrieb erhalten. Für das adaptive Testen ist die Anwendung der Item-Response-Theorie eine Voraussetzung. Die Item-Response-Theorie ist eine Theorie über den Zusammenhang zwischen einer Testaufgabe (Item) und der Lösung der Aufgabe durch eine Person (Response).

Anstelle der vor allem im englischsprachigen Sprachraum gebräuchlichen Bezeichnung Item-Response-Theorie wird im europäischen Sprachraum häufig die Bezeichnung probabilistische Testtheorie verwendet. Probabilistische Modelle nehmen eine stochastische Beziehung zwischen dem Antwortverhalten einer Person, der Fähigkeit der Person und der Aufgabenschwierigkeit an. Diese Beziehung wird durch eine mathematische, nicht lineare logistische Funktion grafisch dargestellt und als Item Response Funktion beziehungsweise als Item Response Curve oder Item Characteristic Curve (ICC) bezeichnet (vgl. Abbildung 1.1).

Abbildung 1.1: Item Characteristic Curve zweier unterschiedlicher Aufgaben, die dem Rasch-Modell (gleicher Anstieg der Kurve) entsprechen



Mit der ICC wird die Wahrscheinlichkeit der richtigen Lösung einer Testaufgabe in Abhängigkeit der Fähigkeit einer Person und der Schwierigkeit einer Aufgabe bestimmt. Es handelt sich also um probabilistische Modelle, in der sich – im Fall des Rasch-Modells[2] – die Wahrscheinlichkeit der richtigen Lösung einer Aufgabe als Funktion

[2] Die verschiedenen logistischen Modelle mit unterschiedlicher Anzahl von Itemparametern werden unter dem Begriff Item Response Theorie zusammengefasst (Rost, 1996, S. 136). Unterschieden werden ein-, zwei- und drei-parametrische Modelle. Bei den einparametrischen Modellen variiert nur die Schwierigkeit (difficulty) zwischen den Items, die Trennschärfe beziehungsweise der Anstieg der Itemfunktion hingegen ist für alle Items gleich. Bei zweiparametrischen Modellen unterscheiden sich die Aufgaben auch in ihrer Trennschärfe (dis-

zweier Modellparameter, der Fähigkeit der antwortenden Person und der Schwierigkeit der betreffenden Aufgabe, ergibt. Je grösser die Fähigkeit ist, desto wahrscheinlicher ist es, dass eine Person eine bestimmte Aufgabe richtig löst. Und je schwieriger eine Aufgabe ist, desto unwahrscheinlicher ist es, dass eine Person mit einer bestimmten Fähigkeit die Aufgabe richtig löst.

Die Beziehung zwischen Lösungswahrscheinlichkeit, Fähigkeit und Aufgabenschwierigkeit (ICC) wird aufgrund der Ergebnisse der Testpersonen für jede Aufgabe bestimmt. Die Kurven geben für Aufgaben mit bestimmten Schwierigkeitsparametern (δ_1 und δ_2 in Abbildung 1.1) die Lösungswahrscheinlichkeit p in Abhängigkeit des Fähigkeits- beziehungsweise des Personenparameters θ an (Kubinger, 2003, S. 3). Ein grosser Vorteil dieser Methode besteht darin, dass sowohl das Testergebnis der Schülerinnen und Schüler (Personenparameter, Fähigkeit) als auch die Schwierigkeit der Aufgaben (Schwierigkeitsparameter, Aufgabenschwierigkeit) auf der gleichen Skala abgebildet werden. Diese Skala ist so konstruiert, dass bei einer Entsprechung von Aufgabenschwierigkeit und Personenfähigkeit die Lösungswahrscheinlichkeit $p = 0.5$ beträgt.

In der Abbildung 1.1 beträgt die Schwierigkeit der Aufgabe 1, für die die Item Characteristic Curve gezeichnet ist, $\delta_1 = 0$. Eine Person, deren Fähigkeit mit $\theta = 0$ geschätzt wird, hat bei dieser Aufgabe eine Lösungswahrscheinlichkeit von $p = 0.5$ beziehungsweise 50 Prozent. Personen, deren Fähigkeit $\theta > 0$ ist, besitzen bei der Aufgabe 1 eine grössere Lösungswahrscheinlichkeit. Personen, deren Fähigkeit $\theta < 0$ ist, besitzen bei der Aufgabe 1 eine kleinere Lösungswahrscheinlichkeit. Die Lokation der Kurve zeigt, ob die Aufgabe eher schwierig oder eher einfach ist. Je weiter links die Kurven liegen, desto einfacher sind die Aufgaben, je weiter rechts die Kurven liegen, desto schwieriger sind die Aufgaben. Die Schwierigkeit der Aufgabe 2, für die ebenfalls die Item Characteristic Curve dargestellt ist, beträgt $\delta_2 = 1$. Für Personen, deren Fähigkeit $\theta = 0$ ist, sinkt die Lösungswahrscheinlichkeit bei der Aufgabe 2 im Vergleich zur Lösungswahrscheinlichkeit bei der Aufgabe 1.

Vorteile probabilistischer Modelle

Die Beziehung zwischen Lösungswahrscheinlichkeit, Fähigkeit einer Person und Schwierigkeit einer Aufgabe ist für die Interpretation der Testergebnisse von entscheidender Bedeutung. Sofern dieses Modell gilt, können Testergebnisse in Bezug zu den Testaufgaben, aber auch und in Bezug zu den in den Interpretationshilfen enthaltenen Umschreibungen der Testaufgaben durch Begriffe und Kompetenzen interpretiert werden. Konkret heisst dies, dass sich aufgrund des individuellen Testergebnisses für jede Aufgabe bestimmen lässt, wie wahrscheinlich es ist, dass ein Schüler oder eine Schülerin die Aufgabe richtig löst. Die Anwendung probabilistischer Testmodelle hat dadurch gegenüber der Anwendung der klassischen Testtheorie den Vorteil, dass die Testergebnisse in Bezug zu Kompetenzen und in diesem Sinne förderorientiert genutzt werden können.

---

crimination), das heisst im Anstieg der Itemfunktion. Im drei-parametrischen Verfahren wird zudem noch die Ratewahrscheinlichkeit parametrisiert (pseudo guessing parameter), die bei Multiple-Choice-Aufgaben unterschiedlich hoch sein kann.

Die Anwendung der Item-Response-Theorie bildet zudem für das computergestützte Testen die eigentliche Grundlage. Wenn das probabilistische Modell gilt, dann hat der Test die Eigenschaft der so genannten spezifischen Objektivität. Das bedeutet, dass die geschätzten Fähigkeitswerte der Person unabhängig von den Schwierigkeitswerten der verwendeten Aufgaben sind und umgekehrt (Leutner, 2001, S. 526). Die Testergebnisse einer Person können unabhängig der Aufgaben berechnet werden, die die Person gelöst hat. Diese Eigenschaft ist eine Voraussetzung dafür, dass ein Test adaptiv verlaufen kann.

2 Kalibrierung der Testaufgaben und Normierung

2.1 Testdesign

Die Kalibrierung der Testaufgaben beziehungsweise die Normierung der Tests von Stellwerk 8 wurde im Frühling 2005 mit allen Schülerinnen und Schülern der 8. Klassen des Kantons St. Gallen durchgeführt. Dass Tests mit einer Population normiert werden, ist eher selten. Populationen sind meist sehr gross, weshalb auf eine Erhebung bei allen Mitgliedern der Population verzichtet und die Normierung mit Hilfe einer repräsentativen Stichprobe aus der Population durchgeführt wird. Weil für Stellwerk 8 sehr viele Testaufgaben normiert werden mussten, war der Einbezug aller Schülerinnen und Schüler der 8. Klassen des Kantons St. Gallen (Referenzpopulation) jedoch sinnvoll. Insgesamt konnten dank der Teilnahme von über 6000 Schülerinnen und Schülern mehr als 1200 Testaufgaben normiert beziehungsweise geeicht werden.

Tabelle 2.1: Verteilung der Aufgaben auf die Schülerinnen und Schüler

Testversionen	Link-Aufgaben	Link-Aufgaben	Kern-Aufgaben
Version 1	A1	A12	B1
Version 2	A2	A1	B2
Version 3	A3	A2	B3
Version 4	A4	A3	B4
Version 5	A5	A4	B5
Version 6	A6	A5	B6
Version 7	A7	A6	B7
Version 8	A8	A7	B8
Version 9	A9	A8	B9
Version 10	A10	A9	B10
Version 11	A11	A10	B11
Version 12	A12	A11	B12

Für die Fachbereiche Mathematik, Deutsch und Englisch wurden je zwölf Testversionen eingesetzt, für Französisch neun und für die Fachbereiche Natur und Technik pro Fach drei Versionen. Die Tests wurden am Computer gelöst, jedoch nicht adaptiv. Die Versi-

onen wurden systematisch auf die Schülerinnen und Schüler verteilt, so dass die Zuordnung der Versionen auf die Schülerinnen und Schüler zufällig war. Innerhalb einer Klasse wurde zwei, maximal drei Schülerinnen und Schülern die gleiche Version zugewiesen. Um Positionseffekte vermeiden zu können, wurden die Aufgaben den Schülerinnen und Schülern nach dem Zufallsprinzip zugewiesen.

Die verschiedenen Testaufgaben wurden pro Fachbereich zu Gruppen zusammengefasst (vgl. Tabelle 2.1). Für die Mathematik wurden beispielsweise 24 Aufgabengruppen gebildet (A1 – A12 und B1 – B12). Die Kernaufgaben (A1 bis A12) wurden jeweils nur auf eine Version verteilt. Die Linkaufgaben (B1 – B12) wurden jeweils auf zwei Testversionen verteilt. Die Versionen waren thematisch gemischt, enthielten folglich Aufgaben von unterschiedlichen Teilbereichen der Mathematik.

2.2 Zusammenstellung der Tests

Damit ein adaptives Testsystem wunschgemäss funktioniert, muss es auf eine genügend grosse Anzahl von Aufgaben zurückgreifen können. Dazu wird eine so genannte Itembank entwickelt. In der Itembank werden erprobte und bewährte Testaufgaben gesammelt, deren Schwierigkeit bekannt ist. Die gegenwärtige Anzahl Aufgaben der Tests von Stellwerk 8 sind in Tabelle 2.2 für jeden geprüften Fachbereich und Teilbereich dargestellt. Die einzelnen Tests werden laufend mit neuen Aufgaben ergänzt.

Tabelle 2.2: Anzahl Aufgaben in der Itembank nach Fachbereich und Teilbereich

Fachbereiche	Teilbereiche	Anzahl Aufgaben
Mathematik		331
	Zahlen, Grössen, Operationen (Arithmetik)	156
	Form und Mass in Ebene und Raum (Geometrie)	86
	Variable, Term, Gleichung (Algebra)	48
	Datendarstellung, Datenanalyse und Zufall, funktionale Zusammenhänge und ihre Darstellungsformen (Stochastik, Funktionen)	41
Deutsch		252
	Hören	83
	Lesen	78
	Sprachreflexion und Rechtschreibung	91
Französisch		152
	Hören	77
	Lesen	75
Englisch		193
	Hören	86
	Lesen	107
Natur und Technik		150
	Biologie	51
	Chemie	49
	Physik	50
TOTAL		1078

Bei der Aufgabenselektion wurde darauf geachtet, dass einerseits die Trennschärfe der Aufgaben genügend hoch ($pt_{bis} > 0.30$) war und die Aufgaben modellkonform waren. Die Trennschärfe entspricht der punkt-biserialen Korrelation zwischen dem Ergebnis bei einer Aufgabe und dem Gesamtergebnis im Test. Die Modellkonformität wurde einerseits für jede Aufgabe überprüft, in dem die Item Characteristic Curve ICC dargestellt wurde. Die Mean-Square-Fit-Statistics (Infit beziehungsweise Outfit-Indices) mussten im Intervall zwischen 0.80 und 1.20 liegen (Wright & Masters, 1982, S. 99). Zudem wurden grafische Modelltests vorgenommen, mit denen überprüft wurde, ob die Schätzung der Schwierigkeitsparameter aus Teilstichproben (Mädchen und Knaben, Schülerinnen und Schüler verschiedener Schultypen sowie Kinder mit unterschiedlicher Erstsprache) zu gleichen Parametern führte.

2.3 Stellwerkskala

Schwierigkeits- und Personenparameter wurden z- standardisiert und anschliessend auf die Stellwerk-Skala mit Mittelwert 500 und Standardabweichung 100 transformiert. Die Testergebnisse der Schülerinnen und Schüler sowie die Schwierigkeitsparameter in der Interpretationshilfe liegen deshalb in der Regel zwischen 200 und 800 Punkten.

Für die Interpretation der Testergebnisse wurde vom Entwicklungsteam des Lehrmittelverlags St. Gallen eine Interpretationshilfe erstellt, in der die Deskriptoren, Begriffe und Beispielaufgaben zu sechs Intervallen von 100 Punkten zusammengefasst sind (200–300 Punkte; 301–400 Punkte; 401–500 Punkte; 501–600 Punkte; 601–700 Punkte; 701–800 Punkte). Jedes Intervall ist durch typische Aufgabenbeispiele illustriert, deren Schwierigkeitsparameter jeweils im entsprechenden Intervall liegen. Eine Aufgabe mit dem Schwierigkeitsparameter von beispielsweise 750 Punkten auf der Stellwerk-Skala wird in der Interpretationshilfe dem Intervall von 700 bis 800 Punkten zugeordnet. Ein Intervall enthält aber auch Kompetenzbeschreibungen beziehungsweise Deskriptoren, die für die Lösung von Aufgaben mit entsprechenden Schwierigkeitsparametern vorausgesetzt werden.

Die Zusammenfassung von Deskriptoren, Begriffen und Beispielaufgaben zu Intervallen erleichtert die Interpretation der Testergebnisse. Das Testergebnis zeigt, über welche Kompetenzen ein Schüler oder eine Schülerin bereits verfügt und welche Kompetenzen angestrebt werden müssen. Dementsprechend sind auch die Informationen für die Lehrpersonen dargestellt, sowohl in einer ausführlichen Dokumentation «*Wie werden die Ergebnisse in den Stellwerk-Tests interpretiert? Von den Testergebnissen zu einer professionellen Beurteilung*.» als auch in einer Kurzversion «*Interpretation der Ergebnisse in Stellwerk-Test.*» (www.stellwerk-check.ch).

Weil die Fähigkeiten mit den Kompetenzbeschreibungen (Deskriptoren, Begriffe und Beispielaufgaben) eines Intervalls umschrieben werden, wurde die Skala mit den Schwierigkeitsparametern so verschoben, dass bei Übereinstimmung der Fähigkeit eines Schülers mit der Aufgabenschwierigkeit – das heisst, Personenparameter und Schwierigkeitsparameter sind identisch – der Schüler die Aufgabe mit einer Wahrscheinlichkeit von $p = 0.62$ (62 Prozent) richtig löst (und nicht mit einer Wahrscheinlichkeit von $p = 0.50$). Durch diese Massnahme konnte vermieden werden, dass die Fähigkeiten von Schülerinnen und Schülern, deren Testergebnisse am unteren Ende eines Intervalls liegen, überschätzt werden.

Die gewählte Lösungswahrscheinlichkeit von p = 0.62 hat zur Folge, dass beispielsweise eine Schülerin, deren Testergebnis am unteren Ende eines Intervalls liegt, eine durchschnittliche Lösungswahrscheinlichkeit der Aufgaben des Intervalls von p = 0.5 hat. Erreicht die Schülerin beispielsweise im Mathematiktest 300 Punkte, dann beträgt die durchschnittliche Lösungswahrscheinlichkeit für alle Aufgaben des Intervalls (zwischen 300 und 400 Punkten) p = 0.5; das heisst, sie löst vermutlich 50 Prozent der Aufgaben dieses Intervalls richtig. Tabelle 2.3 zeigt, wie sich die Lösungswahrscheinlichkeit für Schülerinnen und Schüler in Abhängigkeit ihrer Fähigkeiten (Personenparameter) und der Aufgabenschwierigkeiten (Schwierigkeitsparameter) verändert.

Erreicht ein Schüler beispielsweise im Mathematiktest 300 Punkte auf der Stellwerk-Skala, dann hat er bei einer Aufgabe, deren Schwierigkeit ebenfalls mit 300 Punkten auf der Stellwerk-Skala angegeben ist, eine Lösungswahrscheinlichkeit von 62 Prozent. Löst der Schüler eine einfachere Aufgabe, beispielsweise eine Aufgabe mit einem Schwierigkeitsparameter von 250 Punkten, dann steigt die Lösungswahrscheinlichkeit auf 82 Prozent. Löst der Schüler eine schwierigere Aufgabe, beispielsweise eine Aufgabe mit einem Schwierigkeitsparameter von 350 Punkten, dann sinkt die Lösungswahrscheinlichkeit auf 50 Prozent.

Tabelle 2.3: Lösungswahrscheinlichkeiten in Abhängigkeit von Testergebnissen und Aufgabenschwierigkeiten

	Aufgabenschwierigkeiten (Schwierigkeitsparameter)		
Testergebnisse (Personenparameter)	300	350	400
400	82%	73%	62%
350	73%	62%	50%
300	62%	50%	38%

Das Testergebnis in Form der Punktzahl auf der Stellwerk-Skala zeigt den Schülerinnen und Schülern mit Hilfe der Interpretationshilfe, über welche Kompetenzen und Begriffe sie mit einer mittleren Wahrscheinlichkeit verfügen beziehungsweise welche Aufgaben sie mit einer mittleren Wahrscheinlichkeit lösen können. Die Lösungswahrscheinlichkeit liegt je nach Fähigkeit innerhalb eines Intervalls zwischen 38 und 82 Prozent. Beispielsweise löst ein Schüler mit der Fähigkeit von 400 Punkten die meisten Aufgaben im Intervall zwischen 300 und 400 Punkten ohne Probleme. Kommt der Schüler nur auf 300 Punkte, hat er beim Lösen der schwierigeren Aufgaben meist noch Probleme. Aufgaben eines tieferen Intervalls, beispielsweise jene zwischen 200 und 300 Punkten, kann der Schüler hingegen mit sehr hoher Wahrscheinlichkeit richtig lösen. Aufgaben eines höheren Niveaus, beispielsweise zwischen 400 und 500 Punkten, kann der Schüler erst mit geringer Wahrscheinlichkeit lösen.

3 Ergebnisse der Normierung 2005

Die Prüfung der verschiedenen Tests hat ergeben, dass die Leistungen in den einzelnen Teilbereichen eines Tests eng miteinander zusammenhängen. Das heisst, dass beispielsweise die Mathematikaufgaben unabhängig ihrer Zugehörigkeit zu einem Teilbereich eine mathematische Kompetenz erfassen, die für die Kompetenzen in den verschiedenen Teilbereihen grundlegend ist. Aus diesem Grund wurden bei sämtlichen Tests eindimensionale Modelle und nicht mehrdimensionale Modelle angewendet. Allerdings ist es möglich und auch sinnvoll, bei der Anwendung von eindimensionalen Modellen die Teilbereiche als Subskalen zu behandeln und in das Modell einzuführen. Dadurch wird es möglich sein, die Ergebnisse wie vorgesehen für Teilbereiche auszuweisen.

Damit die Testergebnisse in den Teilbereichen ausgewiesen werden können, wurde der Algorithmus so programmiert, dass die der Fähigkeit des Schülers beziehungsweise der Schülerin entsprechende Aufgabe für Mathematik, Deutsch, Französisch und Englisch jeweils systematisch abwechselnd aus den Teilbereichen gesucht wird, so dass bei Testende mehr oder weniger aus allen Teilbereichen gleich viele Aufgaben bearbeitet wurden. Nachdem die allgemeine fachliche Kompetenz (Gesamtwert, Personenparameter) – beispielsweise mathematische Kompetenz – mit hoher Zuverlässigkeit geschätzt wurde, berechnet das System am Ende ebenfalls anhand des Rasch-Modells die Kompetenzen zu den einzelnen Teilbereichen.

3.1 Mathematik

Von den 421 in der Eichung eingesetzten Testaufgaben werden 331 in der Version 1.0 von Stellwerk eingesetzt. Jede Aufgabe wurde von mindestens 455, zum Teil aber von über 1000 Schülerinnen und Schülern bearbeitet.

Tabelle 3.1 gibt einen Überblick zu den wichtigsten Kennwerten. Zum einen enthält die Tabelle die Korrelationskoeffizienten zur Beschreibung der Zusammenhänge der Teilbereiche. Die Teilbereiche sind in der ersten Spalte mit dem Namen, in der zweiten Spalte mit der dazu gehörigen Nummer beschriftet. Die Nummern der Teilbereiche befinden sich auch in der ersten Zeile (1 bis 7). Die Höhe der Korrelationen sind der Matrix zu entnehmen. Beispielsweise beträgt der Zusammenhang zwischen den Teilbereichen Grössen (2) und Proportionen (5) r = .93. Korrelationen grösser r = .80 gelten als hoch.

In der Mathematik liegen die Reliabilitätskoeffizienten zwischen r = .81 für den Teilbereich Abbildungen und Konstruktionen und r = .85 für die Teilbereiche Operationen und Gleichungen. In der zweituntersten Zeile befindet sich die durchschnittliche Schwierigkeit des Tests. Sie beträgt für die Mathematik insgesamt 48%. Das heisst, in Durchschnitt sind die Aufgaben von rund 48 Prozent der Schülerinnen und Schüler gelöst worden. Die unterste Zeile enthält die durchschnittliche Trennschärfe der Testaufgaben. Mit pt_{bis} = 0.44 ist dieser Wert für die Mathematik hoch. In der letzten Spalte rechts der Tabelle 3.1 ist die Anzahl Aufgaben pro Teilbereich enthalten. Der Teilbereich Gleichungen umfasst 48 Aufgaben.

Tabelle 3.1: Angaben zum Mathematiktest, Version 1.0

	TB	1	2	3	4	5	6	7	Anzahl Aufgaben
Zahlen und Zahlenbereiche	1								52
Grössen	2	.92							53
Operationen	3	.92	.91						51
Gleichungen	4	.92	.91	.95					48
Proportionen	5	.92	.93	.90	.92				41
Abbildungen und Konstruktionen	6	.89	.87	.92	.92	.90			40
Geometrische Berechungen	7	.90	.90	.92	.93	.90	.93		46
		1	2	3	4	5	6	7	Total
Reliabilität		.82	.82	.85	.85	.82	.81	.83	.87
Schwierigkeit		53%	45%	53%	46%	50%	47%	38%	48%
Trennschärfe		0.42	0.43	0.47	0.48	0.41	0.39	0.49	0.44

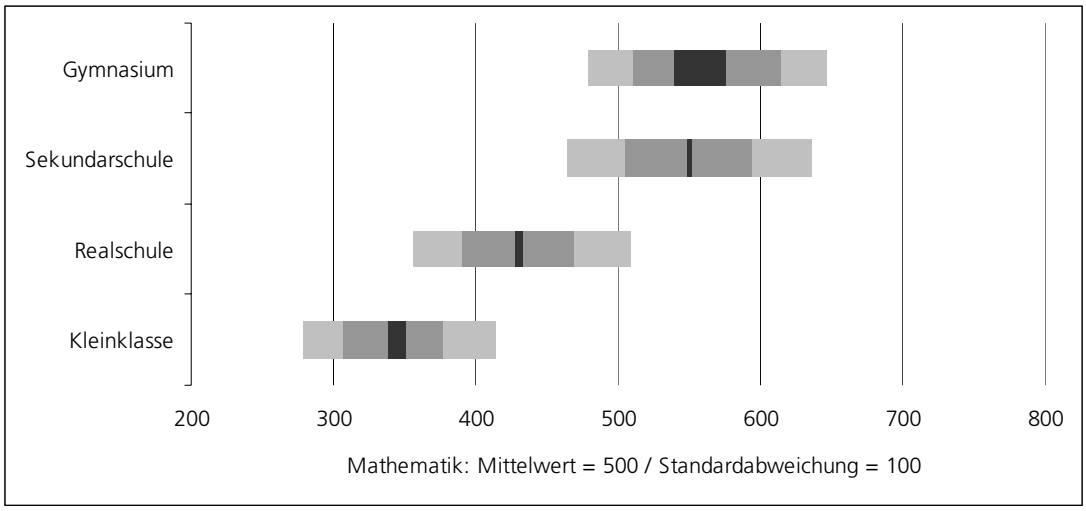
Die Ausweisung der Testergebnisse für sieben Teilbereiche hat sich beim ersten Durchgang von Stellwerk nicht bewährt. Für die zuverlässige Schätzung der mathematischen Kompetenz mussten die Schülerinnen und Schüler rund 40 Aufgaben lösen. Dies hatte zur Folge, dass die Berechnung der Testergebnisse in den Teilbereichen aufgrund von fünf bis siegen Aufgaben erfolgen musste. Weil das Lösen der Testaufgaben relativ viel Zeit in Anspruch nimmt, konnte der Test nicht verlängert werden. Aus diesem Grund wurden die Mathematikaufgaben zu vier Teilbereichen zusammengefasst. Die Kennwerte sind in der Tabelle 3.2 enthalten.

Tabelle 3.2: Angaben zum Mathematiktest: revidierte Version 1.1

	TB	1	2	3	4	Anzahl Aufgaben
Arithmetik	1					156
Geometrie	2	.93				86
Algebra	3	.96	.92			48
Stochastik, Funktionen	4	.95	.90	.92		41
		1	2	3	4	Total
Reliabilität		.84	.84	.85	.85	.87
Schwierigkeit		50%	46%	50%	42%	48%
Trennschärfe		0.44	0.48	0.41	0.44	0.44

Anhand der Daten der Normierung der Stellwerk-Tests wurden die mathematischen Kompetenzen für die teilnehmenden Schülerinnen und Schüler berechnet. Abbildung 3.1 zeigt die Ergebnisse nach Schultypen. Der kleine schwarze Balken in der Mitte gibt an, in welchem Bereich der wahre Mittelwert statistisch gesichert liegt. Die dunkelgrau schattierten Balken links und rechts vom Mittelwert geben den Bereich an, in dem die mittleren 50 Prozent der Leistungen liegen. Zählt man noch die hellgrau schattierten Balken links und rechts der dunkelgrauen dazu, so erhält man den Bereich, in dem 90 Prozent der Leistungen der Schülerinnen und Schüler liegen.

Abbildung 3.1: Mathematikleistungen nach Schultyp



Anmerkung: Gymnasium: n = 40 Schüler/innen, M = 558 Punkte, SD = 59 Punkte
Sekundarschule: n = 3659 Schüler/innen, M = 551 Punkte, SD = 68 Punkte
Realschule n = 2094 Schüler/innen, M = 430 Punkte, SD = 60 Punkte
Kleinklasse n = 267 Schüler/innen, M = 355 Punkte, SD = 55 Punkte

Es gilt zu beachten, dass das Ergebnis der Gymnasien nur von wenigen Schülerinnen und Schülern zustande gekommen ist. Die Differenzen zwischen den Ergebnissen der Schülerinnen und Schüler der verschiedenen Schultypen sind gross. Allerdings zeigt die Abbildung auch, dass die Überschneidungsbereiche zwischen den Schülerinnen und Schülern verschiedener Schultypen relativ gross sind. Rund ein Viertel der Schülerinnen und Schüler der Realschule erreichen gleich gute oder bessere Leistungen als die schlechtesten Schülerinnen und Schülern der Sekundarschule. Noch grösser sind die Überschneidungsbereiche zwischen den Schülerinnen und Schüler der Kleinklasse und der Realschule. Die Abbildung zeigt, dass Stellwerk auch zu einer gerechteren Beurteilung der Schülerinnen und Schüler genutzt werden kann; eine Beurteilung, bei der die Kompetenzen und nicht der Schulabschluss beziehungsweise der Schultyp im Vordergrund steht.

3.2 Deutsch

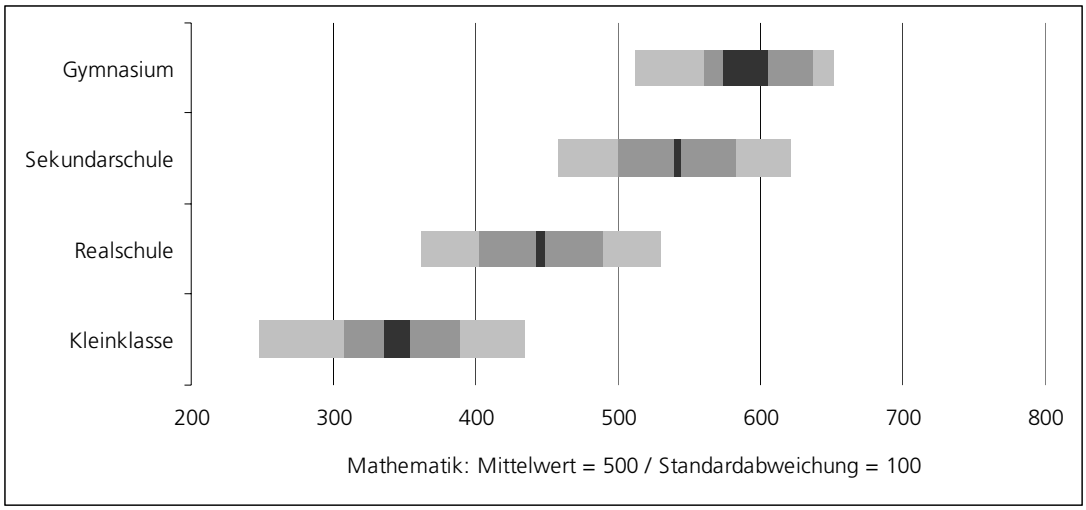
Von den 298 in der Eichung eingesetzten Testaufgaben werden 252 in der Version 1.0 von Stellwerk eingesetzt. Jede Aufgabe wurde von mindestens 479, zum Teil aber von über 1000 Schülerinnen und Schülern bearbeitet.

Tabelle 3.3: Angaben zum Deutschtest

	Hören	Lesen	Sprachreflexion und Rechtschreibung	Anzahl Aufgaben
Hören				83
Lesen	0.89			78
Sprachreflexion und Rechtschreibung	0.89	0.87		91
	Hören	Lesen	Sprachreflexion und Rechtschreibung	Total
Reliabilität	0.67	0.68	0.71	0.75
Schwierigkeit	66%	64%	58%	62%
Trennschärfe	0.33	0.33	0.39	0.35

Tabelle 3.3 gibt einen Überblick zu den wichtigsten Kennwerten. Sie enthält die gleichen Angaben wie Tabelle 1, wobei nun sämtliche Teilbereiche mit Namen und ohne Nummern bezeichnet sind. Die Höhe der Korrelationen zwischen den Teilbereichen sind etwas geringer als in der Mathematik, aber ebenfalls sehr hoch. Die Testaufgaben sind insgesamt deutlich einfacher als jene der Mathematik. Die durchschnittliche Schwierigkeit beträgt 62 Prozent richtig gelöste Aufgaben. Die mittlerer Trennschärfe liegt bei pt_{bis} = .35. Die Reliabilitäten liegen zwischen r = 0.67 und r = 0.75. Der Reliabilitätskoeffizient hängt unter anderem von der Anzahl Aufgaben ab. Dies ist mit ein Grund, weshalb die Reliabilitäten beim Mathematiktest höher liegen.

Abbildung 3.2: Deutschleistungen nach Schultyp.



Anmerkung: Gymnasium: n = 45 Schüler/innen, M = 590 Punkte, SD = 53 Punkte
Sekundarschule: n = 3616 Schüler/innen, M = 542 Punkte, SD = 63 Punkte
Realschule n = 2069 Schüler/innen, M = 446 Punkte, SD = 66 Punkte
Kleinklasse n = 269 Schüler/innen, M = 344 Punkte, SD = 74 Punkte

Abbildung 3.2 zeigt die Verteilung der Deutschleistungen nach Schultyp. Im Gegensatz zur Mathematik, liegt der Mittelwert der Schülerinnen und Schüler der Gymnasien statistisch signifikant über dem Mittelwert der Schülerinnen und Schüler der Sekundarschulen.

Die Überschneidungsbereiche verschiedener Schultypen sind in Deutsch wesentlich grösser als in der Mathematik. Einzelne Schülerinnen und Schüler der Realschulen erreichen in Deutsch sogar gleich gute oder bessere Leistungen wie die schwächsten Schülerinnen und Schüler der Gymnasien. Es gilt aber auch hier wieder zu berücksichtigen, dass das Ergebnis der Gymnasien nur gerade von 45 Schülerinnen und Schülern stammt.

3.3 Französisch

Von den 176 in der Eichung eingesetzten Testaufgaben werden 152 in der Version 1.0 von Stellwerk eingesetzt. Jede Aufgabe wurde von mindestens 709, zum Teil aber von über 1515 Schülerinnen und Schülern bearbeitet. Tabelle 3.4 gibt einen Überblick zu den wichtigsten Kennwerten. Die Höhe der Korrelation zwischen den Teilbereichen Hören und Lesen ist sehr hoch. Dies ist dadurch zu erklären, dass es beim Hörverstehen wie beim Lesenverstehen um die Kompetenz geht, die Sprache zu verstehen beziehungsweise den Sinn eines Textes bei unterschiedlicher Vorgabe zu erfassen.

Tabelle 3.4: Angaben zum Deutschtest

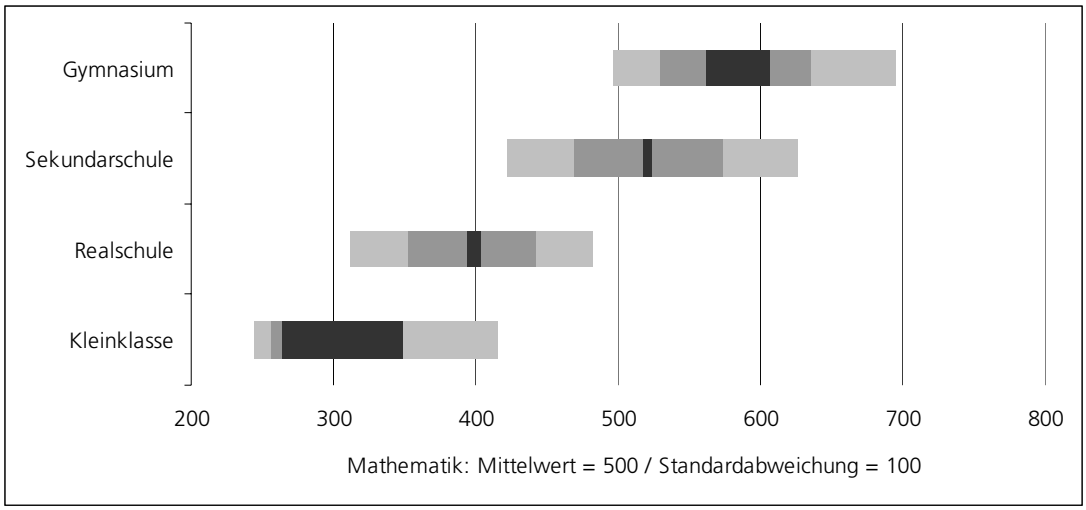
	Hören	Lesen	Anzahl Aufgaben
Hören			77
Lesen	0.95		75
Reliabilität	Hören 0.78	Lesen .77	Total .78
Schwierigkeit	65%	65%	65%
Trennschärfe	0.36	0.36	0.36

Die Testaufgaben sind insgesamt deutlich einfacher als jene der Mathematik, aber ähnlich schwierig wie die Testaufgaben für das Fach Deutsch. Die durchschnittliche Schwierigkeit beträgt 65 Prozent richtig gelöste Aufgaben. Die mittlere Trennschärfe liegt bei $pt_{bis} = .36$. Die Reliabilitäten liegen bei $r = 0.78$.

Abbildung 3.3 zeigt die Verteilung der Leistungen in Französisch nach Schultyp. Die Ergebnisse der Kleinklassen stammen von nur sieben Schülerinnen und Schülern. Wie in Deutsch liegt der Mittelwert der Schülerinnen und Schüler der Gymnasien statistisch signifikant über dem Mittelwert der Schülerinnen und Schüler der Sekundarschulen. Die Überschneidungsbereiche verschiedener Schultypen sind in Französisch eher kleiner als in Deutsch. Die besten Schülerinnen und Schüler der Realschulen erreichen im Fran-

zösischtest aber gleich gute oder bessere Leistungen wie die schwächsten Schülerinnen und Schüler der Sekundarschulen.

Abbildung 3.3: Leistungen in Französisch nach Schultyp



Anmerkung: Gymnasium: n = 46 Schüler/innen, M = 585 Punkte, SD = 76 Punkte
Sekundarschule: n = 3651 Schüler/innen, M = 521 Punkte, SD = 79 Punkte
Realschule n = 769 Schüler/innen, M = 398 Punkte, SD = 69 Punkte
Kleinklasse n = 7 Schüler/innen, M = 306 Punkte, SD = 56 Punkte

3.4 Englisch

Von den 236 in der Eichung eingesetzten Testaufgaben werden 193 in der Version 1.0 von Stellwerk eingesetzt. Jede Aufgabe wurde von mindestens 487, zum Teil aber von bis zu 1495 Schülerinnen und Schülern bearbeitet. Tabelle 3.5 gibt einen Überblick zu den wichtigsten Kennwerten. Die Höhe der Korrelation zwischen den Teilbereichen Hören und Lesen ist sehr hoch. Dies ist dadurch zu erklären, dass es beim Hörverstehen wie beim Leseverstehen um die Kompetenz geht, die Sprache zu verstehen beziehungsweise den Sinn eines Textes bei unterschiedlicher Vorgabe zu erfassen.

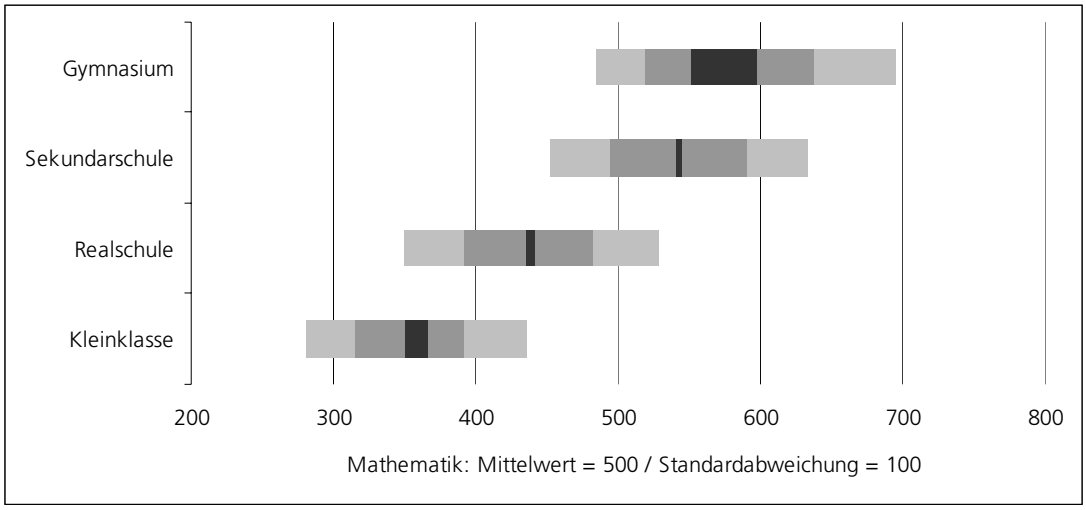
Tabelle 3.5: Angaben zum Deutschtest

	Hören	Lesen	Anzahl Aufgaben
Hören			86
Lesen	0.95		107
Reliabilität	Hören 0.74	Lesen .75	Total .79
Schwierigkeit	57%	65%	62%
Trennschärfe	0.41	0.41	0.41

Die Testaufgaben sind insgesamt deutlich einfacher als jene der Mathematik, aber ähnlich schwierig wie die jene des Deutschtests. Die durchschnittliche Schwierigkeit beträgt 65 Prozent richtig gelöste Aufgaben. Die mittlerer Trennschärfe liegt bei pt_{bis} = .36. Die Reliabilitäten liegen zwischen r = 0.74 und r = .79.

Abbildung 3.4 zeigt die Verteilung der Leistungen in Englisch nach Schultyp. Die Ergebnisse nach Schultypen liegen etwas näher beieinander als in Deutsch und Französisch, was sich beispielsweise bei den Schülerinnen und Schülern der Gymnasien und der Sekundarschulen zeigt. Die Überschneidungsbereiche der Leistungen der Schülerinnen und Schüler verschiedener Schultypen sind in Englisch grösser als in der Mathematik und in Französisch. Einzelne Schülerinnen und Schüler der Realschulen erreichen – gleich wie in Deutsch – auch in Englisch zum Teil gleich gute oder bessere Leistungen als die schwächsten Schülerinnen und Schüler der Gymnasien. Es gilt aber auch hier wieder zu berücksichtigen, dass das Ergebnis der Gymnasien nur gerade von 45 Schülerinnen und Schülern stammen.

Abbildung 3.4: Leistungen in Französisch nach Schultyp.



Anmerkung:

Gymnasium:	n = 45 Schüler/innen,	M = 574 Punkte,	SD = 79 Punkte
Sekundarschule:	n = 3663 Schüler/innen,	M = 542 Punkte,	SD = 70 Punkte
Realschule	n = 2092 Schüler/innen,	M = 438 Punkte,	SD = 69 Punkte
Kleinklasse	n = 205 Schüler/innen,	M = 359 Punkte,	SD = 61 Punkte

3.5 Natur und Technik

Für den Fachbereich Natur und Technik wurden drei unabhängige Tests entwickelt. Die Schülerinnen und Schüler einer Klasse lösten deshalb entweder den Biologietest, den Chemietest oder den Physiktest. Die Tests konnten deshalb nur unabhängig voneinander skaliert werden. Wie stark die Ergebnisse in Biologie, Chemie und Physik zusammenhängen, lässt sich nicht bestimmen.

In der Biologie werden von den 78 in der Eichung eingesetzten Testaufgaben 51 in der Version 1.0 von Stellwerk eingesetzt, in der Chemie werden von ursprünglich 65 Aufgaben 49 eingesetzt, in der Physik von den ursprünglich 66 Aufgaben 50. Die Aufga-

ben wurden von mindestens 580, zum Teil von bis zu 1516 Schülerinnen und Schülern bearbeitet.

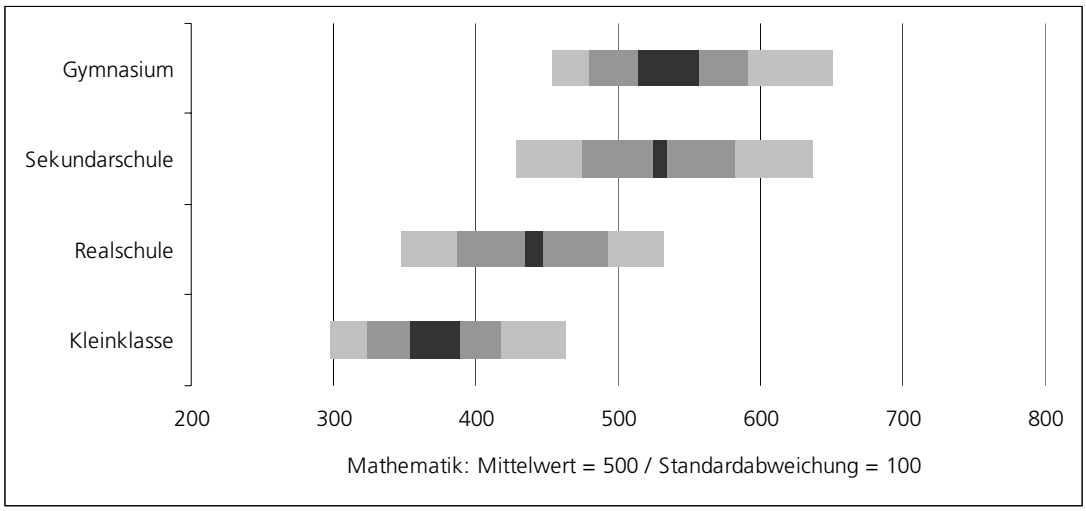
Tabelle 3.6 gibt einen Überblick zu den wichtigsten Kennwerten, wobei wie erwähnt die Berechnung der Korrelationen zwischen den Tests nicht möglich ist. Die Tests zum Bereich Natur und Technik sind deutlich schwieriger als die Sprachtests. Die durchschnittliche Schwierigkeit liegt in der Physik bei 41 Prozent richtig gelöster Aufgaben. Die mittlerer Trennschärfe liegt für den Biologietest bei pt_{bis} = 0.35, für den Chemietest bei pt_{bis} = 0.37 und für den Physiktest bei pt_{bis} = 0.41.

Tabelle 3.6: Angaben zu den Tests im Fachbereich Natur und Technik

	Biologie	Chemie	Physik
Anzahl Aufgaben	51	49	50
Reliabilität	0.26	0.30	0.27
Schwierigkeit	53%	45%	41%
Trennschärfe	0.35	0.37	0.41

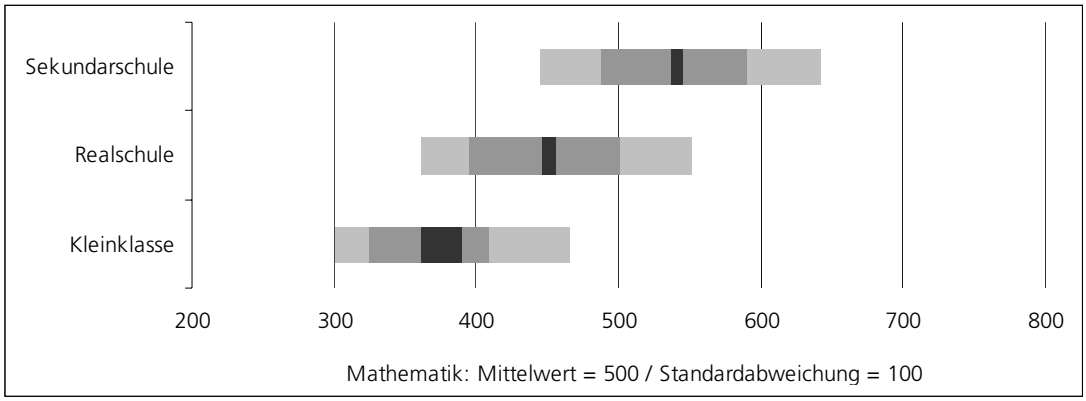
Die Abbildung 3.5 bis Abbildung 3.7 zeigen die Verteilung der Leistungen in den drei Tests des Fachbereichs Natur und Technik sowie die dazugehörigen Mittelwerte und Standardabweichungen.

Abbildung 3.5: Leistungen im Biologietest nach Schultyp



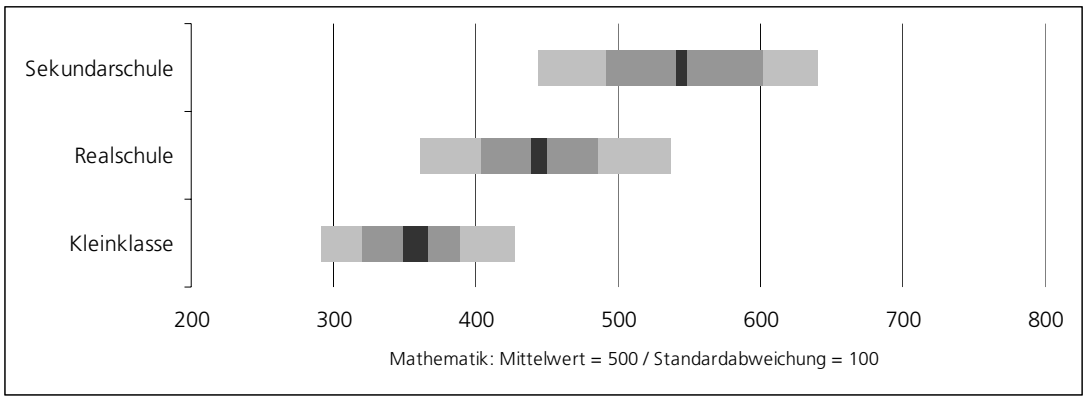
Anmerkung: Gymnasium: n = 44 Schüler/innen, M = 536 Punkte, SD = 72 Punkte
Sekundarschule: n = 1160 Schüler/innen, M = 530 Punkte, SD = 81 Punkte
Realschule n = 527 Schüler/innen, M = 441 Punkte, SD = 75 Punkte
Kleinklasse n = 46 Schüler/innen, M = 371 Punkte, SD = 60 Punkte

Abbildung 3.6: Leistungen im Chemietest nach Schultyp



Anmerkung: Sekundarschule: n = 1294 Schüler/innen, M = 541 Punkte, SD = 76 Punkte
Realschule n = 889 Schüler/innen, M = 451 Punkte, SD = 75 Punkte
Kleinklasse n = 84 Schüler/innen, M = 375 Punkte, SD = 66 Punkte

Abbildung 3.7: Leistungen im Physiktest nach Schultyp



Anmerkung: Sekundarschule: n = 1128 Schüler/innen, M = 545 Punkte, SD = 78 Punkte
Realschule n = 617 Schüler/innen, M = 445 Punkte, SD = 68 Punkte
Kleinklasse n = 119 Schüler/innen, M = 357 Punkte, SD = 50 Punkte

4 Funktionsweise des adaptiven Testens im Jahr 2006

4.1 Testergebnisse im Jahr 2006

Sämtliche Skalen der Stellwerk-Tests wurden auf einen Mittelwert von M = 500 und eine Standardabweichung von SD = 100 normiert. Das bedeutet, dass der Mittelwert der Referenzpopulation (Alle Schülerinnen und Schüler der 8. Klasse des Kantons St. Gallen) im Jahr 2005 M = 500 Punkte beträgt. Diese Normierung erfolgte allerdings nicht aufgrund eines adaptiven Tests. Bei der Normierung im Jahr 2005 wurden die Testaufgaben den Schülerinnen und Schülern in Form von bis zu 24 unterschiedlichen Versionen pro Fachbereich zur vollständigen Bearbeitung vorgelegt.

Im Jahr 2006 lösten wiederum alle Schülerinnen und Schüler der 8. Klasse des Kantons St. Gallen die Stellwerk-Tests, allerdings als computergestützte adaptive Tests. Die Ergebnisse in Form der Mittelwerte und der Standardabweichungen sind in der Tabelle 4.1 enthalten. Die Mittelwerte liegen sehr nahe bei jenen der Eichstichprobe. Für die Fachbereiche Deutsch, Englisch und Chemie beträgt der Unterschied nur gerade 1 Punkt, in der Physik liegt die Differenz bei 3 Punkten, in der Mathematik bei 5 Punkten. Etwas grösser sind die Abweichungen im Fachbereiche Biologie (11 Punkte) und im Fachbereich Französisch (18 Punkte).

Tabelle 4.1: Ergebnisse im Jahr 2006

	Mittelwert	Standard-abweichung	Anzahl Schüler/innen
Deutsch	499	98	6255
Mathematik	505	138	6270
Englisch	499	113	6106
Französisch	518	101	4507
Chemie	499	90	3607
Physik	503	113	3812
Biologie	489	102	4955

Die Unterschiede zwischen den Ergebnissen bei der Normierung im Jahr 2005 und bei der adaptiven Durchführung im Jahr 2006 sind sehr ähnlich. Die geringen Unterschiede können dahin gehend interpretiert werden, dass der Algorithmus korrekt funktioniert. Es ist nicht zu erwarten, dass sich die Leistungen der Schülerinnen und Schüler eines Kantons innerhalb eines Jahres stark ändern, auch wenn es sich um zwei verschiedene Kohorten handelt und ein geringer Kohorteneffekt durchaus möglicht ist.

4.2 Häufigkeit der Aufgabenzuteilung durch den Algorithmus und Schwierigkeitsparameter der Aufgaben

Eine weitere Möglichkeit, die Funktionsweise des Algorithmus zu überprüfen, besteht darin, die Aufgaben nach der Häufigkeit, nach der sie im Testsystem eingesetzt wurden, darzustellen. Rund 68 Prozent der Schülerinnen und Schüler erreichen Testergebnisse, die zwischen 400 und 600 Punkten liegen, rund 95 Prozent erreichen Testergebnisse, die zwischen 300 und 700 Punkten liegen. Aufgrund dieser Verteilung müssen Aufgaben im mittleren Schwierigkeitsbereich häufig, solche in den Extrembereichen eher selten gewählt werden.

Abbildung 4.1: Häufigkeit der Aufgabenwahl nach Schwierigkeitsparameter der Aufgaben: Mathematik

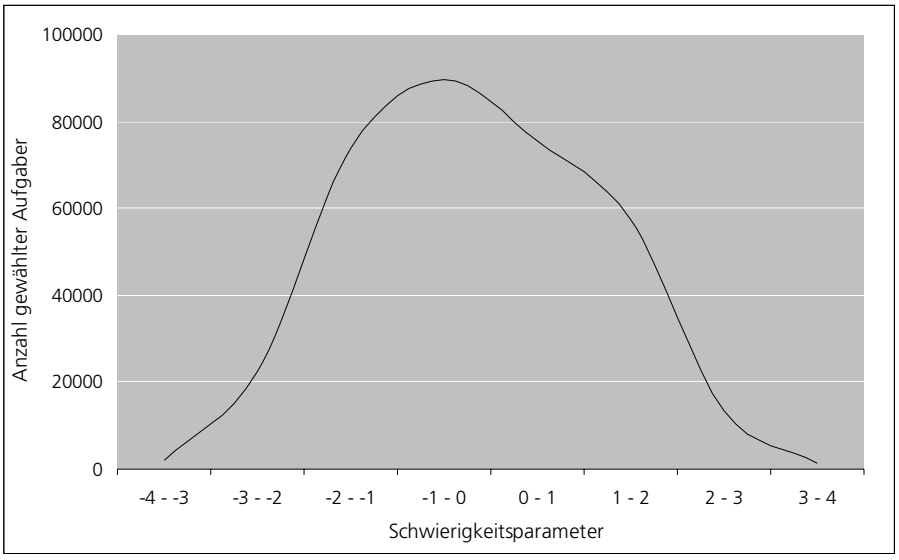
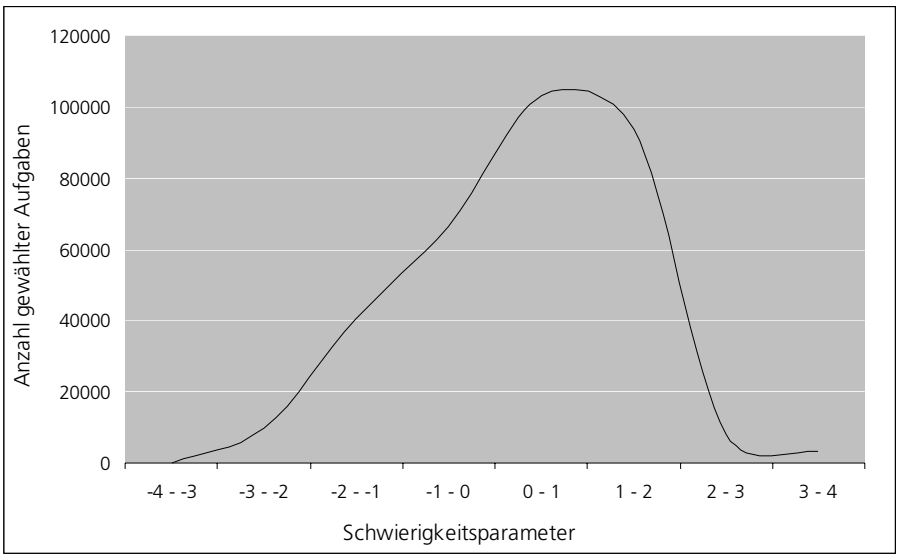


Abbildung 4.2: Häufigkeit der Aufgabenwahl nach Schwierigkeitsparameter der Aufgaben: Deutsch



Die Abbildungen 4.1 und 4.2 zeigen, dass das System entsprechend der Verteilung der Personenparameter Aufgaben mit mittleren Schwierigkeitsparametern häufiger wählt als sehr schwierige und sehr einfache Aufgaben.

5 Ausblick

An testtheoretischen Kriterien beurteilt, funktioniert das Testsystem Stellwerk wie vorgesehen. Ein adaptives Testsystem muss allerdings ständig weiterentwickelt werden. Aus diesem Grund wird beim Durchgang 2007 ein weiterer Algorithmus erprobt, der die Selbstkalibrierung von neuen Testaufgaben zulässt. Sobald ein normierter Aufgabenpool besteht, wird es aufgrund der Anwendung der Item-Response-Theorie möglich, die Schwierigkeitsparameter innerhalb des Testsystems zu berechnen. Analog der Schätzung des Personenparameters innerhalb des adaptiven Testens wird bei der Selbstkalibrierung eine neue Testaufgabe so lange erprobt, bis es nur noch zu geringen Schwankungen bei der Schätzung des Schwierigkeitsparameters kommt. Sobald die Selbstkalibrierung abgeschlossen ist, kann eine neue Testaufgabe im Test eingesetzt werden. Die weiteren Erfahrungen mit dem Testsystem Stellwerk werden in einem Bericht voraussichtlich Ende 2007 dokumentiert.

6 Literatur

Amelang, M. & Schmidt-Atzert (2006). *Psychologische Diagnostik und Intervention*. Heidelberg: Springer.

Drasgow, F. & Olson-Buchanan, J.B. (1999). *Innovations in Computerized Assessment.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kubinger, K. (2002). Adaptives Testen. In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 1–9). Weinheim, Basel, Berlin: Beltz.

Leutner, D. (2001). Pädagogisch-psychologische Diagnostik. In D.H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 521–530). Weinheim, Basel, Berlin: Beltz.

Lienert, G. (1967). *Testaufbau und Testanalyse*. Weinheim, Berlin, Basel: Beltz.

Rost, J. (2003). *Lehrbuch Testtheorie – Testkonstruktion*. Bern, Göttingen, Toronto, Seattle: Huber.

Van der Lindern, W.J. & Glas, C.A.W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, Boston, London: Kluwer.

Wainer, H. (2000). *Computerized Adaptive Testing. A Primer.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wright, B. & Masters, G.N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.