



Universität Zürich  
Institut für Bildungsevaluation

**Institut für Bildungsevaluation**  
Assoziiertes Institut  
der Universität Zürich

Standardisierte Erfassung der sprachlichen Kompetenzen  
im Fachbereich «Texte schreiben»  
Kurzbericht zuhanden des Erziehungsrates des Kantons St. Gallen

Urs Moser  
Zürich, 6. Juni 2008

## **Inhalt**

1	Ausgangslage .....	3
2	Durchführung.....	3
3	Beurteilung der Texte.....	5
4	Beurteilungszuverlässigkeit.....	8
5	Berechnung der Testergebnisse .....	9
6	Ergebnisse.....	12
6.1	Ergebnisse nach Geschlecht .....	12
6.2	Ergebnisse nach Schultypen .....	12
6.3	Ergebnisse nach Klassen .....	13
7	Fazit.....	14

## 1 Ausgangslage

Im Kanton St. Gallen wird das Testsystem *Stellwerk* im 8. und 9. Schuljahr flächendeckend eingesetzt. Stellwerk umfasst Tests für die Fachbereiche Deutsch, Englisch, Französisch, Mathematik sowie Natur und Technik. Es handelt sich um adaptive Tests, die sich den Fähigkeiten der Schülerinnen und Schüler anpassen und ausschliesslich am Computer gelöst werden. Adaptive Tests haben zur Folge, dass die Schülerinnen und Schüler in der Regel mit Aufgaben getestet werden, die weder viel zu schwierig noch viel zu einfach sind. Die Auswahl der Testaufgaben wird durch einen Algorithmus gesteuert. Während die Schülerinnen und Schüler die Testaufgaben bearbeiten, werden vom Testsystem fortwährend ihre Fähigkeiten geschätzt. Sobald sich in der Schätzung der Fähigkeiten keine grossen Änderungen mehr abzeichnen, wird der Test abgebrochen und das definitive Testergebnis festgehalten.

Adaptive Testsysteme haben gegenüber traditionellen «Papier-und-Bleistift-Tests» den Vorteil, dass die Objektivität bei der Testdurchführung gesichert ist – sofern keine technischen Probleme auftreten – und dass der Computer bei der Korrektur keine Fehler macht. Die Korrektur des Computers hat allerdings auch Nachteile. Computergestützte Tests prüfen vorwiegend reproduktive Fähigkeiten, weil ausführliche Antworten auf offene Fragen oder Texte vom Computer meist nicht in der gewünschten Art und Weise korrigiert und bewertet werden können. Produktive Fähigkeiten wie Schreiben und Sprechen können deshalb am Computer nicht getestet werden.

Das Institut für Bildungsevaluation der Universität Zürich führt seit mehreren Jahren Tests zur Erfassung der Schreibfähigkeiten durch. Im Frühjahr 2008 wurden im Auftrag des Kantons St. Gallen zum ersten Mal sämtliche Schülerinnen und Schüler des 8. Schuljahres im Fachbereich «Texte schreiben» getestet.

## 2 Durchführung

Die Durchführung des Schreibenlasses wurde auf den Monat Februar 2008 festgesetzt. Weil während dieser Jahreszeit in einzelnen Gemeinden Ferien sind oder Schulen Skilager eingeplant haben, wurden drei zweitägige Testzeitfenster zur Auswahl gestellt: 14./15. Februar, 21./22. Februar und 28./29. Februar.

Zu jedem Testzeitfenster wurden jeweils zwei Themen mit Schreibaufträgen vorgelegt, aus denen die Schülerinnen und Schüler ein Thema auswählen mussten. Tabelle 1 zeigt, welche Themen von den Schülerinnen und Schülern wie häufig gewählt wurden. Beispielsweise haben zum ersten Zeitpunkt 64 Prozent der Schülerinnen und Schüler einen Text zum Thema «Striktes Handyverbot an den Urdorfer Schulen» gewählt, während sich 36 Prozent für das Thema «Kampf gegen die Kilos schon im frühen Kindesalter» entschieden haben.

Die Themen wurden den Schülerinnen und Schülern mit kurzen Texten, die auf Zeitungsartikeln basieren, vorgestellt. Im Anschluss an den kurzen Text wurden jeweils zwei Aufgaben zum Inhalt des Zeitungsartikels gestellt. Mit der ersten Aufgabe wurde von den Schülerinnen und Schülern ein argumentativer Text verlangt. Beispielsweise mussten die Fragen «Was spricht deiner Meinung nach für ein Handyverbot an allen Schweizer Schulen? Was spricht deiner Meinung nach dagegen?» beantwortet werden. Die zweite Aufgabe

verlangte eine Beschreibung eigener Beobachtungen und Erfahrungen mit dem Thema. Beispielsweise musste die Frage «Welche Beobachtungen und Erfahrungen machst du zu diesem Thema in deinem Alltag?» beantwortet werden.

Tabelle 1: Prozentanteil gewählter Themen nach Testzeitfenster

Thema	Anzahl Schülerinnen und Schüler	Prozentanteil pro Thema
Striktes Handyverbot an den Urdorfer Schulen	1600	64%
Kampf gegen die Kilos schon im frühen Kindesalter	916	36%
Schuluniform	866	60%
Gefährliches Ski- und Snowboardfahren	587	40%
Anonyme Bewerbung für Lehrstellen möglich	422	27%
Spucken und Abfallentsorgung in der Öffentlichkeit	1136	73%

Im Sinne einer standardisierten schriftlichen Anleitung wurden die Schülerinnen und Schüler zu folgendem Vorgehen aufgefordert:

---

Soll ein Handyverbot an allen Schweizer Schulen eingeführt werden?

*Bearbeite zu diesem Thema die beiden Aufgaben auf den folgenden Seiten.*

*Du gehst wie folgt vor:*

- *Lies die zwei Aufgaben zuerst durch.*
- *Schreibe dann die Texte zu den zwei Aufgaben auf ein Notizpapier.*
- *Korrigiere den Entwurf.*
- *Achte auf die Rechtschreibung und schreibe so, dass deine Texte gut lesbar sind.*
- *Schreibe danach deine Texte zu den zwei Aufgaben auf die ausgeteilten Blätter.*

---

Im Anschluss an die Durchführung in den Schulklassen mussten die Texte der Schülerinnen und Schüler von den Lehrpersonen kopiert und die Originale dem Institut für Bildungsevaluation zur Korrektur und Beurteilung zugestellt werden. Die Ergebnisse in Form einer Punktzahl wurden anschliessend von der Firma «Cybersystems», die im Projekt «Stellwerk» für die Informatik zuständig ist, den Lehrpersonen auf dem Internet zur Verfügung gestellt. Ab Montag, 28. April 2008, waren die Ergebnisse im Bereich «Texte schreiben» für die Lehrpersonen beziehungsweise für die Schülerinnen und Schüler auf dem Internet einsehbar.

Insgesamt verfassten 5528 Schülerinnen und Schüler einen Text. Ein Text wurde allerdings zu spät eingeschickt und ist bei den folgenden Auswertungen nicht berücksichtigt.

### 3 Beurteilung der Texte

Zur Beurteilung der Texte wurde ein Kriterienraster entsprechend bisheriger Erfahrungen mit der Korrektur von Texten und auf der Grundlage der Testtheorie entwickelt. Das Kriterienraster wurde in Anlehnung an das Zürcher Textanalyseraster von Nussbaumer & Sieber (1994)<sup>1</sup> entwickelt, wobei aus testtheoretischen Überlegungen nur ein Teil dieser Kriterien berücksichtigt und in adaptierter Form eingesetzt wurde. Das Beurteilungsverfahren entspricht einem analytischen Vorgehen, bei dem verschiedene Aspekte eines Textes nach verbal formulierten Abstufungen bewertet werden (Analytical Scoring)<sup>2</sup>.

Beurteilt wurde zum einen, wie gut die zwei kommunikativen Aufgaben erfüllt wurden (Tabelle 2, Seite 6). Dazu wurden die aufgeführten Argumente für oder gegen die jeweilige Regelung – beispielsweise Tragen einer Schuluniform – gezählt (Aufgabe 1) und bewertet, ob Beobachtungen oder Erfahrungen aus dem Alltag im Text erwähnt wurden (Aufgabe 2). Zudem wurde die Textlänge erfasst. Die drei Kriterien zur Beurteilung der kommunikativen Fähigkeiten sind *quantitativ* abgestuft.

Zum anderen wurden anhand von formalen Kriterien linguistische Kompetenzen beurteilt (Tabelle 2, Seite 7). Das Kriterium «Argumente verbinden» wurde nur anhand des Textes zu Aufgabe 1, das Kriterium «Textaufbau» nur anhand des Textes zu Aufgabe 2 angewendet. Die übrigen formalen Kriterien (Verständlichkeit, Wortwahl, Rechtschreibung, Satzzeichen, Satzverbindungen, Syntax sowie Ästhetisches Wagnis und Kreativität) wurden auf den gesamten Text zu beiden Aufgaben angewendet. Die neun formalen Kriterien zur Beurteilung der linguistischen Fähigkeiten sind *qualitativ* abgestuft.

Die Kriterien wurden anhand von drei, vier oder fünf Kategorien umschrieben. Das heisst, dass entweder 0, 1 oder 2 Punkte (drei Kategorien), 0, 1, 2 oder 3 Punkte (vier Kategorien) oder 0, 1, 2, 3 oder 4 Punkte (fünf Kategorien) vergeben wurden. Insgesamt konnten 31 Punkte erreicht werden.

Die zwölf Kriterien wurden gleich wie zwölf Testaufgaben behandelt und einer Itemanalyse nach der klassischen Testtheorie unterzogen. Tabelle 1 enthält die statistischen Kennwerte der inhaltlichen Kriterien.

In der ersten Spalte der Tabelle 1 befindet sich die Bezeichnung des Beurteilungskriteriums, in der zweiten Spalte die quantitative Abstufung zur Beurteilung des Textes anhand des Kriteriums. Beispielsweise wurden die kommunikativen Fähigkeiten anhand der Häufigkeit von erwähnten Argumenten beurteilt.

In der dritten Spalte der Tabelle 1 befindet sich der Prozentanteil der Texte, denen die Ausprägung des Kriteriums zugeordnet wurde. In 43 Prozent der Texte waren beispielsweise bei der ersten Aufgabe vier oder mehr als vier Argumente erwähnt. In 35 Prozent der Texte waren drei Argumente erwähnt, in 18 Prozent zwei Argumente und in 4 Prozent war nur ein Argument erwähnt.

---

<sup>1</sup> Nussbaumer, M. & Sieber, P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In P. Sieber (Hrsg.), *Sprachfähigkeiten – besser als ihr Ruf und nötiger denn je! Ergebnisse aus einem Forschungsprojekt* (S. 141–186). Aarau: Sauerländer.

<sup>2</sup> Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

In der vierten Spalte der Tabelle 1 sind die Angaben zur Trennschärfe des Kriteriums enthalten. Der Trennschärfekoeffizient zeigt bei einem Test, inwiefern eine Aufgabe Schülerinnen und Schüler mit hohem Gesamtwert von Schülerinnen und Schülern mit niedrigem Gesamtwert trennt. Angewendet auf die Beurteilung von Texten zeigt die Trennschärfe, wie gut die Punktzahl eines Kriteriums mit der Gesamtbeurteilung übereinstimmt. Ein hoher Trennschärfekoeffizient zeigt, dass gute Texte anhand des Kriteriums positiv und schlechte Texte eher negativ beurteilt werden. Ein niedriger Trennschärfekoeffizient (um 0) besagt, dass gute und schlechte Texte anhand des Kriteriums gleich oder ähnlich beurteilt werden, und ein negativer Koeffizient bedeutet, dass gute Texte anhand des Kriteriums oft negativ, schlechte oft positiv beurteilt werden. Der Trennschärfekoeffizient sollte nicht kleiner als  $r_{it} = 0.30$  sein.

Die Reliabilität beziehungsweise die Messgenauigkeit erreicht ein Cronbach-Alpha von  $\alpha = .78$ , was darauf hinweist, dass die Beurteilungskriterien ziemlich konsistent angewendet wurden und sich relativ gut eigneten, zuverlässig zwischen guten und weniger guten Texten zu unterscheiden<sup>3</sup>.

Tabelle 2: Kriterien zur Beurteilung der kommunikativen Fähigkeiten

Beurteilungskriterium	Quantitative Abstufungen	Prozentanteil Texte	Trennschärfe
<i>Aufgabe 1: Argumentation</i>			
– Argumente erwähnt	0 Argument aufgeführt	0%	0.50
	1 Argument aufgeführt	4%	
	2 Argumente aufgeführt	18%	
	3 Argumente aufgeführt	35%	
	4 oder mehr Argumente aufgeführt	43%	
<i>Aufgabe 2: Beobachtungen</i>			
– Beobachtungen/Erfahrungen	keine Beobachtung/Erfahrung	0%	0.51
	1 Beobachtung/Erfahrung	10%	
	2 bis 3 Beobachtungen/Erfahrungen	58%	
	mehr als drei Beobachtungen/Erfahrungen	32%	
<i>Aufgaben 1 und 2</i>			
– Textlänge (insgesamt)	weniger als eine halbe Seite	0%	0.70
	eine halbe Seite	3%	
	eine Seite	22%	
	eineinhalb Seiten	37%	
	mehr als eineinhalb Seiten	38%	

Tabelle 3 enthält die gleichen Angaben wie Tabelle 2, allerdings zu den Kriterien, die zur Beurteilung der linguistischen Kompetenz angewendet wurden.

<sup>3</sup> Bei einmaliger Testvorgabe wird zur Berechnung der Reliabilität der Koeffizient «Cronbach-Alpha» verwendet. Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse*. Basel: Beltz.

Tabelle 3: Kriterien zur Beurteilung der linguistischen Fähigkeiten

Beurteilungskriterium	Qualitative Abstufungen	Prozentanteil Texte	Trennschärfe
<i>Aufgabe 1</i>			
– Argumente verbinden	Argumente aneinandergereiht	12%	0.55
	Argumente verknüpft, kommentiert	70%	
	logischer Aufbau der Argumentation	18%	
<i>Aufgabe 2</i>			
– Textaufbau	kein Aufbau erkennbar	8%	0.66
	Aufbau erkennbar	45%	
	klarer Aufbau	35%	
	klarer Aufbau, in sich abgeschlossener Text	12%	
<i>Aufgaben 1 und 2</i>			
– Verständlichkeit	nicht verständlich: unklare Aussagen	1%	0.44
	grösstenteils verständlich	17%	
	gut verständlich	82%	
– Wortwahl	einfach, simpel	6%	0.55
	adäquat	85%	
	elaboriert, herausragend, überraschend	9%	
– Rechtschreibung (inkl. Fallfehler)	Rechtschreibung hindert die Lesbarkeit	7%	0.52
	trotz Fehlern gut lesbar	57%	
	(nahezu) fehlerfrei	36%	
– Satzzeichen	rudimentär vorhanden (Punkt, kein Komma)	15%	0.55
	meist korrekte Satzzeichensetzung	52%	
	(nahezu) fehlerfrei	33%	
– Satzverbindung (Konjunktionen)	keine oder einfache Sätze	4%	0.57
	immer gleich	57%	
	abwechslungsreich	39%	
– Syntax	teilweise korrekte Sätze	4%	0.56
	einfache korrekte Sätze	52%	
	komplexe korrekte Sätze (HS und NS)	44%	
– Ästhetisches Wagnis/Kreativität	wagt wenig, einfache Lösung	4%	0.69
	wagt etwas, Kreativität erkennbar	61%	
	wagt viel, kreativ	29%	
	ausgesprochen kreativer Text, unerwartete Ideen	6%	

Am strengsten beurteilt (beziehungsweise eher selten erreichten die Schülerinnen und Schüler bei diesen Kriterien die höchste Punktzahl) wurden das «Ästhetisches Wagnis» und der «Textaufbau». Diese Kriterien waren für die Schülerinnen und Schüler schwierig zu erfüllen. Am mildesten beurteilt beziehungsweise eher gut erfüllt wurden das Kriterium «Verständlichkeit» sowie die kommunikativen Kriterien (Argumente aufführen, Meinung anbringen, eigene Beobachtungen und Erfahrungen darlegen).

## 4 Beurteilungszuverlässigkeit

Die Texte wurden von vier Lehrpersonen und Studierenden der Germanistik (Rater) nach den vorgegebenen Kriterien korrigiert und beurteilt. Zwei der vier Rater arbeiten bereits seit mehreren Jahren am Institut für Bildungsevaluation bei der Korrektur von Texten mit und sind im Korrigieren von solchen Tests versiert. Zwei der Rater nahmen zum ersten Mal an einer Korrektur von Schreiben anlässen teil. Das Korrekturteam wurde bewusst klein gehalten, sodass die Standardisierung der Beurteilung dank gemeinsamer Absprache und regelmässiger Kontrolle hoch gehalten werden konnte.

In einer ersten Schulungsphase wurde der Kriterienkatalog auf seine Tauglichkeit überprüft. Anhand einer repräsentativen Auswahl von Texten wurde zudem ein gemeinsamer Beurteilungsmassstab gesucht. Im Anschluss an diese Phase wurden zwanzig Texte doppelt korrigiert und Abweichungen bei der Beurteilung diskutiert. Die Überprüfung der unabhängigen Beurteilung der gleichen Texte führte zu keinen grossen Differenzen zwischen den beurteilenden Personen.

Während der gesamten Korrekturphase wurden rund zwanzig Prozent der Texte doppelt korrigiert, um die Beurteilungsübereinstimmung ständig zu überprüfen. Abweichungen in der Beurteilung wurden laufend diskutiert mit dem Ziel, die Beurteilungsübereinstimmung (Inter-Rater-Reliabilität) hoch zu halten. Tabelle 4 enthält Informationen zur Beurteilungsübereinstimmung bei doppelt korrigierten Texten. Die Angaben entsprechen den Durchschnittswerten aller Doppelkorrekturen, die aus sämtlichen möglichen Kombinationen erhalten wurden. Die doppelt korrigierten Texte wurden ausgewogen nach Themen auf die Rater verteilt. In der zweiten Spalte ist pro Kriterium angegeben, wie hoch die vollständige Übereinstimmung in Prozent ist.

Tabelle 4: Prozentuale Übereinstimmung und Inter-Rater-Reliabilität

Beurteilungskriterium	Vollständige Übereinstimmung	Kappa
Argumente erwähnt	57%	0.88
Beobachtungen und Erfahrungen	61%	0.70
Textlänge (insgesamt)	96%	0.85
Argumente verbinden	63%	0.22
Textaufbau	51%	0.49
Verständlichkeit	76%	0.32
Wortwahl	84%	0.41
Rechtschreibung	63%	0.45
Satzzeichen	66%	0.53
Satzverbindungen (Konjunktionen)	52%	0.22
Syntax	55%	0.19
Ästhetisches Wagnis/Kreativität	54%	0.41

Die Übereinstimmung ist bei jenen Kriterien am höchsten, die anhand von quantitativen Abstufungen beurteilt werden konnten (beispielsweise Testlänge). Bei Kriterien, die anhand von wenigen qualitativen Abstufungen beurteilt wurden, ist die Übereinstimmung in der Regel deutlich geringer.

Indem die Anzahl der Übereinstimmungen berechnet und am Anteil der zufälligen Übereinstimmung relativiert wird, kann auch das statistische Zusammenhangsmass «Kappa» zur Bestimmung der Beurteilungsübereinstimmung berechnet werden (Tabelle 4, dritte Spalte)<sup>4</sup>. Der Kappakoeffizient kann Werte zwischen  $-1$  und  $+1$  annehmen. Der maximale Wert wird bei totaler Übereinstimmung erreicht. Bei einer systematisch gegensätzlichen Einstufung wird der Wert negativ. Das Kappa hängt unter anderem auch von der Anzahl Abstufungen eines Kriteriums ab. Liegt das Kappa  $> 0.70$ , dann wird die Übereinstimmung als gut bezeichnet.

Bei verschiedenen Kriterien liegt das Kappa unter  $0.70$ . Dabei muss allerdings beachtet werden, dass bei sehr ungleichen Randverteilungen bereits wenige Abweichungen in der Beurteilung das Kappa sehr klein werden lassen. Ein Blick auf die vollständige Übereinstimmung zeigt, dass das Kappa relativ schnell tief wird. Damit sich die unterschiedlichen Beurteilungsmassstäbe für die Schülerinnen und Schüler nicht negativ auswirken, müssen bestimmte Verzerrungen allerdings bei der Berechnung der Testergebnisse berücksichtigt werden.

## 5 Berechnung der Testergebnisse

Dass der gleiche Text trotz vorgegebener Kriterien, Schulungsphase und ständiger Kontrolle von mehreren Personen nicht immer gleich beurteilt wird, ist aufgrund des Interpretationsspielraums bei offen gestellten Aufgaben zu erwarten. Wie kann aber verhindert werden, dass systematische Unterschiede bei der Beurteilung von Texten keine negativen Auswirkungen auf die Ergebnisse der Schülerinnen und Schüler haben?

Unterschiedliche Beurteilungsmassstäbe können mit verschiedenen schwierigen Testaufgaben verglichen werden: Je strenger ein Kriterium von einer beurteilenden Person (Rater) angewendet wird, desto schwieriger ist die Aufgabe für die Schülerinnen und Schüler. Beurteilt beispielsweise Person A systematisch strenger als Person B, dann ist dies natürlich für all jene Schülerinnen und Schüler ungerecht, deren Text von Person A beurteilt wird. Wird die Strenge oder Milde in der Beurteilung der Texte bei der Berechnung der Testergebnisse nicht berücksichtigt, dann ist beispielsweise der gleiche Schreibanlass je nach beurteilender Person entweder etwas einfacher oder etwas schwieriger.

Bei der Beurteilung eines Textes bestimmen mindestens vier Faktoren das Testergebnis: (1) *Die Fähigkeit der Schülerin oder des Schülers*: Leistungsstärkere Schülerinnen und Schüler erhalten eine höhere Beurteilung als leistungsschwächere. (2) *Die Schwierigkeit des Kriteriums (Item)*: Ein Kriterium ist dann schwierig, wenn die Schülerinnen und Schüler

---

<sup>4</sup> Zuerst wird der Anteil der beobachteten Übereinstimmungen  $P_0$  berechnet (Diagonale in einer  $k \times k$ -Felder-Tafel). Danach wird aufgrund der Zeilen- und Spaltensummen der Anteil aller zufälligen Übereinstimmungen  $P_e$  bestimmt. Kappa entspricht der Differenz zwischen  $P_0 - P_e$  über  $1 - P_e$ . [Bortz, J. (1993). *Statistik für Sozialwissenschaftler*. Berlin: Springer. (S. 538)].

bei der Anwendung des Kriteriums generell eher niedrige Beurteilungen erhalten. Dies trifft beispielsweise für das Kriterium «Textaufbau» zu. (3) *Die Strenge oder Milde der beurteilenden Person (Rater)*: Die Kriterien werden von den Ratern jeweils nicht exakt gleich interpretiert. (4) *Das Thema (Task)*: Texte zu spezifischen Themen werden nicht immer gleich streng beurteilt.

Um Effekte der beurteilenden Person und des gewählten Themas auf die Berechnung der Testergebnisse im Bereich «Texte schreiben» statistisch kontrollieren zu können, werden diese beiden Faktoren als Facetten der Urteilsituation aufgefasst und bei der Berechnung der Ergebnisse berücksichtigt<sup>5</sup>. Mit der Anwendung der Item-Response-Theorie ist es möglich, die Beurteilungsstrenge der beurteilenden Personen sowie das Thema ins Testmodell einzubeziehen und bei der Berechnung der Testergebnisse entsprechend zu berücksichtigen. Ein solches Vorgehen wird auch als «Multi-Faceted Measurement» oder als «Multi-Facetten-Modell» bezeichnet<sup>6</sup>. Die beurteilenden Personen (Rater) und die Themen (Task) werden als Facetten eines mehrdimensionalen Testmodells betrachtet, sodass sich die mangelnde Beurteilungsübereinstimmung nicht negativ auf das Ergebnis der Schülerinnen und Schüler auswirkt<sup>7</sup>.

Damit diese Modelle angewendet werden können, ist eine möglichst grosse Anzahl doppelter Korrekturen notwendig. Insgesamt wurden mehr als 1000 Texte doppelt korrigiert. Dabei wurde darauf geachtet, dass pro Text und Rater ungefähr die gleiche Anzahl Texte doppelt korrigiert wurde. Dies führte dazu, dass jeder Rater 250 Texte doppelt korrigierte und von jedem Thema rund 170 Texte doppelt korrigiert wurden.

Die Analyse zeigt (Anhang 1 und 2), dass Rater 3 deutlich strenger beurteilt als Rater 1 und 2. Rater 4 liegt mit seiner Beurteilung näher bei Rater 1 und 2, die sehr ähnlich beurteilen. Die grösste Distanz zwischen zwei Ratern beträgt 0.35 Logits. Bei den Themen (Task) sind die Unterschiede noch etwas grösser. Die Themen «Schuluniform» und «Gefährliches Ski- und Snowboardfahren» wurden am strengsten beurteilt. Etwas milder wurden die Themen «Striktes Handyverbot an den Urdorfer Schulen» und «Spucken und Abfallentsorgung in der Öffentlichkeit» beurteilt und am mildesten wurden die Themen «Kampf gegen die Kilos schon im frühen Kindesalter» und «Anonyme Bewerbung für Lehrstellen möglich» beurteilt. Die grösste Differenz zwischen dem Thema «Kampf gegen die Kilos schon im frühen Kindesalter» und «Schuluniform» betrug 0.57 Logits.

Die Anwendung der Item-Response-Theorie (Multi-Facetten-Modell) bei der Berechnung der Ergebnisse führte dazu, dass die Testrohwerte (Anzahl Punkte) in die standardisierte Normalverteilung transformiert werden. Dabei wurden die Testrohwerte so transformiert, dass – analog der Stellwerksskala – der Mittelwert 500 Punkte und die Standardabweichung 100 Punkte betragen (vgl. Abbildung 1). Diese Skala hat die Eigenschaft, dass rund 68 Prozent der Ergebnisse zwischen 400 und 600 Punkten liegen, rund 95 Prozent zwischen 300 und 700 Punkten und nahezu alle Ergebnisse zwischen 200 und 800 Punkten.

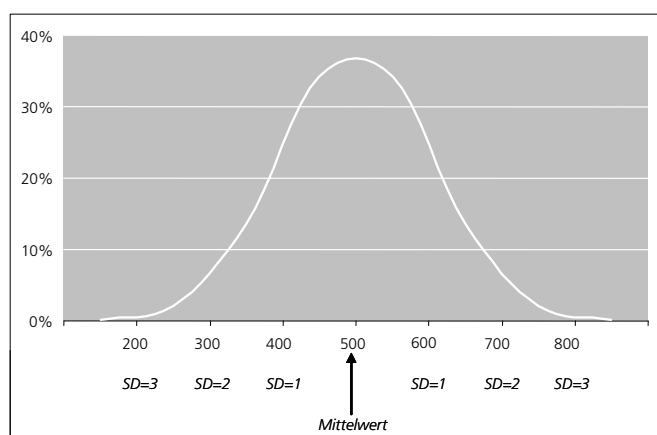
---

<sup>5</sup> Eckes, T. (2004). Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In A. Wolff, T. Ostermann & C. Chlosta (Hrsg.), *Integration durch Sprache* (S. 485–518). Regensburg: Fachverband Deutsch als Fremdsprache.

<sup>6</sup> McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

<sup>7</sup> Rost, J. (2003). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.

Abbildung 1: Verteilung der Testergebnisse



Die Anzahl Punkte zeigt den Schülerinnen und Schülern, wie gut sie innerhalb der Vergleichsgruppe – 5527 Schülerinnen und Schüler, die den Text geschrieben haben – abgeschnitten haben.

Die Anwendung der Item-Response-Theorie hat auch den Vorteil, dass sich die Schwierigkeiten der Kriterien und die Fähigkeiten der Schülerinnen und Schüler auf derselben Skala beziehungsweise mit demselben Massstab abbilden lassen. Zwischen den Fähigkeiten der Schülerinnen und Schüler und den Beurteilungskriterien wird eine Beziehung hergestellt. Tabelle 5 zeigt zusammenfassend, welche Schreibkompetenzen innerhalb eines bestimmten Intervalls vorhanden sind.

Tabelle 5: Kompetenzbeschreibungen nach Punkteintervallen

Punkteintervall	Kompetenzbeschreibungen
200 bis 300 Punkte	Die Texte sind sehr kurz und der kommunikative Auftrag (Argumente oder Beobachtungen erwähnen) ist minimal erfüllt. Die Texte sind nur teilweise verständlich, auch weil sie viele Rechtschreibfehler enthalten.
301 bis 400 Punkte	Die Texte umfassen in der Regel zwischen einer halben und einer ganzen Seite. Der kommunikative Auftrag ist erfüllt, wobei meist nur ein Argument und eine Beobachtung enthalten sind. Die Texte sind grösstenteils verständlich.
401 bis 500 Punkte	Die Texte umfassen jeweils mindestens eine Seite. Es werden einfache, syntaktisch korrekte Sätze angewendet, wobei die Satzverbindungen immer gleich sind. Die Wortwahl ist dem Thema angepasst und die Texte sind trotz Rechtschreibfehlern gut lesbar. Ein Textaufbau ist erkennbar und die Argumente sind verknüpft.
501 bis 600 Punkte	Die Texte umfassen jeweils eineinhalb Seiten und erfüllen den kommunikativen Auftrag umfassend mit mehreren Argumenten und Beobachtungen. Ein klarer Aufbau ist erkennbar. Die Satzzeichen sind meist korrekt und die Texte sind nahezu fehlerfrei. Die Satzverbindungen sind abwechslungsreich und die Texte sind kreativ (Ästhetisches Wagnis).
601 bis 700 Punkte	Die Texte umfassen mehr als eineinhalb Seiten, folgen einem logischen Aufbau und sind in sich abgeschlossen. Die Wortwahl ist herausragend und überraschend und die Texte überzeugen durch ausgesprochene Kreativität und unerwartete Ideen.
701 bis 800 Punkte	Dito (keine zusätzliche Beurteilungskriterien nachweisbar)

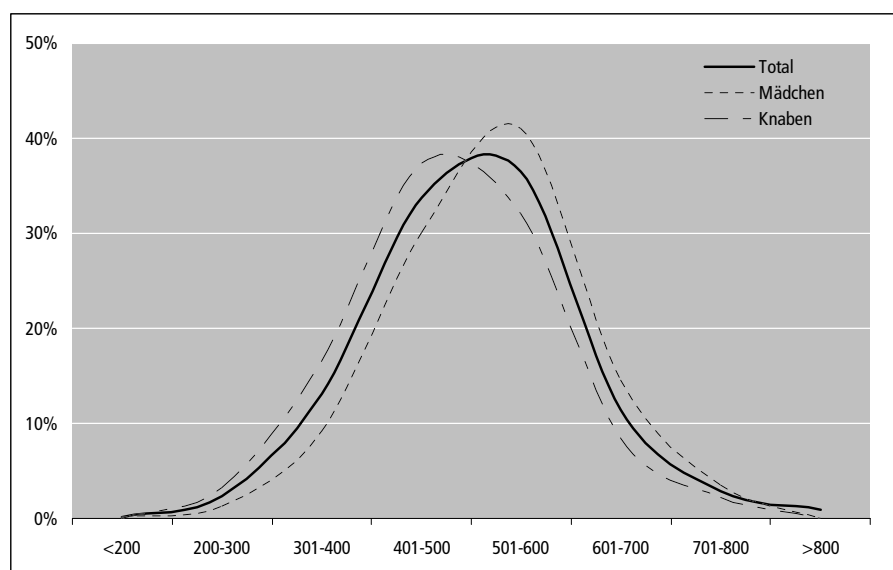
Die Intervalle sind hierarchisch aufgebaut. Das bedeutet für die Interpretation der Ergebnisse, dass Schülerinnen und Schüler, die ein bestimmtes Intervall erreichen (beispielsweise 501 bis 600 Punkte), nicht nur die Fähigkeiten des Intervalls 501 bis 600 Punkte vorweisen, sondern auch über alle Fähigkeiten der darunterliegenden Intervalle verfügen. Wenn beispielsweise ein Text mit 650 Punkten beurteilt wird, dann gilt für diesen Text selbstverständlich auch, dass der kommunikative Auftrag erfüllt ist.

## 6 Ergebnisse

### 6.1 Ergebnisse nach Geschlecht

Abbildung 2 zeigt die Verteilung der Ergebnisse nach Geschlecht. Der Mittelwert der Mädchen liegt bei 519 Punkten, der Mittelwert der Knaben bei 482 Punkten. Die Differenz von 37 Punkten ist statistisch signifikant und von mittlerer Bedeutung. Die Geschlechterdifferenzen sind innerhalb der Schultypen ähnlich gross. In der Kleinklasse beträgt die Differenz 27 Punkte, in der Realschule 34 Punkte und in der Sekundarschule 33 Punkte.

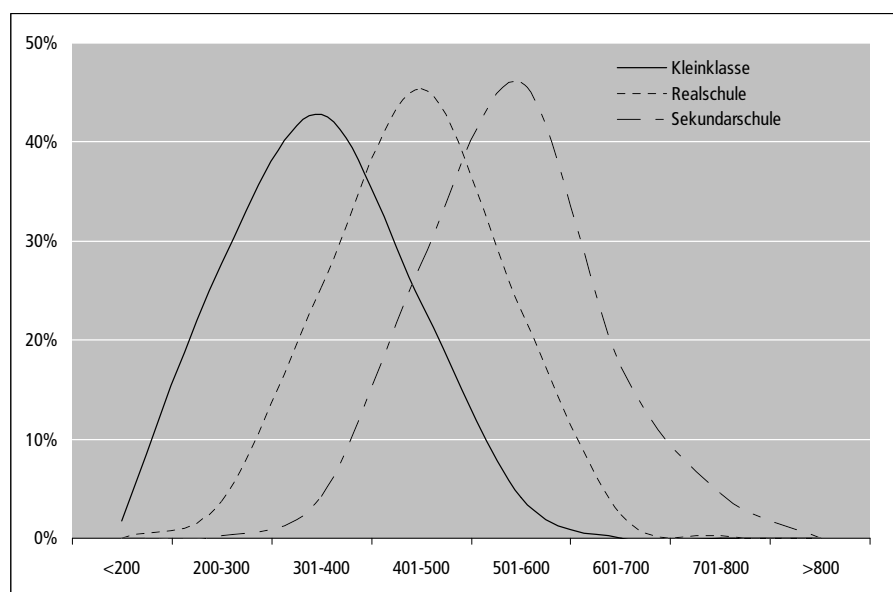
Abbildung 2: Ergebnisse nach Geschlecht



### 6.2 Ergebnisse nach Schultypen

Abbildung 3 zeigt die Ergebnisse nach den Schultypen. Die Verteilungskurven entsprechen den Erwartungen. Die Texte der Schülerinnen und Schüler der Sekundarschulen wurden am häufigsten mit 500 bis 600 Punkten beurteilt (Mittelwert = 539 Punkte). Die Texte der Schülerinnen und Schüler der Realschule wurden am häufigsten mit 400 bis 500 Punkten beurteilt (Mittelwert = 447 Punkte). Die Texte der Schülerinnen und Schüler der Kleinklasse wurden am häufigsten mit 300 bis 400 Punkten beurteilt (Mittelwert = 354 Punkte).

Abbildung 3: Dreiteilige Sekundarschule nach Abteilungen



Die Verteilungskurven zeigen, welche Vorteile eine schultypenunabhängige Beurteilung dank eines objektiven Verfahrens für einzelne Schülerinnen und Schüler haben kann. Rund 40 Prozent der Schülerinnen und Schüler der Kleinklassen erreichten mit ihren Texten eine Beurteilung, die über dem Mittelwert der Realschule liegt. Rund ein Drittel der Schülerinnen und Schüler der Realschule erreichten mit ihren Texten eine Beurteilung, die über dem Mittelwert der Sekundarschule liegt.

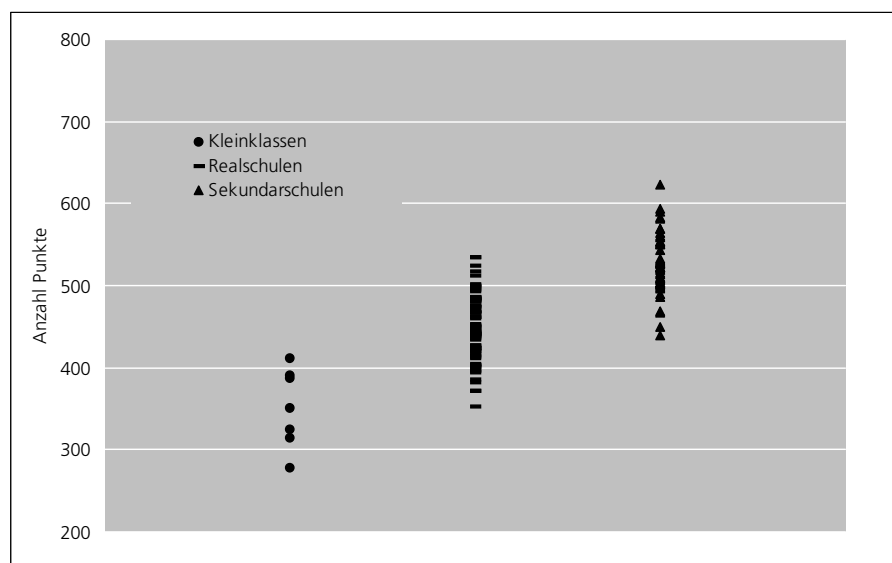
Insgesamt verfassten auch 19 Schülerinnen und Schüler des Gymnasiums einen Text. Der Mittelwert dieser Gruppe liegt mit 578 Punkten noch 39 Punkte über dem Mittelwert der Schülerinnen und Schüler der Sekundarschule.

### 6.3 Ergebnisse nach Klassen

Abbildung 4 zeigt die Ergebnisse der beteiligten Klassen nach Schultypen. Die individuellen Testergebnisse wurden zu einem Klassenmittelwert aggregiert und sind in der Abbildung als Kreis (Kleinklasse), Querstrich (Realschulklasse) oder Dreieck (Sekundarschulklasse) dargestellt. Mittelwerte wurden nur dann gebildet, wenn von einer Kleinklasse mehr als acht, von einer Real- oder Sekundarschulklasse mehr als zehn Texte vorlagen.

Die Differenzen zwischen den Mittelwerten der Sekundarschulklassen sind sehr gross und betragen knapp 200 Punkte. Der tiefste Mittelwert liegt bei 425 Punkten, der höchste bei 623 Punkten. Ähnlich gross sind die Unterschiede zwischen den Ergebnissen der Realschulklassen. Sie streuen zwischen 353 und 535 Punkten. Viele Realschulklassen erreichen deutlich bessere Klassenergebnisse als Sekundarschulklassen. Die Mittelwerte der Kleinklassen mit mindestens acht Schülerinnen und Schülern liegen zwischen 278 und 410 Punkten.

Abbildung 4: Klassenmittelwerte nach Abteilung



## 7 Fazit

Die Durchführung eines flächendeckenden Schreibanlasses bei den Schülerinnen und Schülern der 8. Klassen verlief ohne organisatorische oder technische Probleme. Der postalische Ablauf könnte aber noch verbessert werden. Die Texte sollten immer pro Klasse und Lehrperson identifizierbar sein, was nicht immer eindeutig der Fall war. Einige Lehrpersonen unterliessen es zudem, den Schülerinnen und Schülern mitzuteilen, dass nur ein Thema zu bearbeiten ist.

Die jeweils zwei zur Wahl stehenden Themen verlangten von den Schülerinnen und Schülern einerseits die Fähigkeit, sich zu einem Thema argumentativ zu äussern sowie ihre eigene Meinung anzubringen und zu begründen. Andererseits mussten sie eigene Beobachtungen und Erfahrungen zum gewählten Thema beschreiben. Beurteilt wurden sowohl kommunikative Kompetenzen (Wurde der Auftrag erfüllt?) als auch linguistische Kompetenzen (Wie wurde der Auftrag erfüllt?).

Die Beurteilungsübereinstimmung der vier Rater (Lehrpersonen und Studierende der Germanistik der Universität Zürich) kann insgesamt als genügend beurteilt werden. Die unabhängige Korrektur gleicher Texte durch zwei Personen führte zu einer durchschnittlichen Inter-Rater-Reliabilität von 0.47 (Kappakoeffizient). Dieser Koeffizient ist etwas zu tief, sollte aber nicht unabhängig von der prozentualen Übereinstimmung beurteilt werden, die mit 65 Prozent ansprechend hoch ist. Im Durchschnitt wurde demnach bei rund zwei Drittel der zu beurteilenden Kriterien die gleiche Punktzahl vergeben.

Bei Kriterien, mit denen die inhaltliche Vollständigkeit anhand zählbarer Textstellen überprüft wurde, erreichten die Rater die höchste Beurteilungsübereinstimmung. Etwas schlechter war die Beurteilungsübereinstimmung, wenn die Rater die Texte aufgrund von formalen Kriterien beurteilten. Die Kriterien zur Beurteilung linguistischer Fähigkeiten

werden deshalb überarbeitet und entsprechend angepasst. Vor allem werden neu bei den meisten Kriterien vier anstelle von drei Kategorien angewendet.

Eher grösser als erwartet war die Bedeutung der unterschiedlichen Themen, zu denen den Schülerinnen und Schülern Aufgaben gestellt wurden. Nicht nur die beurteilenden Personen, sondern vor allem auch die gestellten Aufgaben spielen für das Testergebnis eine bedeutende Rolle.

Dank der Anwendung der Item-Response-Theorie konnten sowohl die unterschiedlichen Beurteilungsmassstäbe der Rater als auch die unterschiedlich beurteilten Aufgaben bei der Berechnung der Testergebnisse der Schülerinnen und Schüler rechnerisch berücksichtigt werden, sodass eine zuverlässige und faire Beurteilung möglich wurde.

Die Ergebnisrückmeldung auf der transformierten Skala (Mittelwert = 500 Punkte und Standardabweichung = 100 Punkte) darf nicht darüber hinwegtäuschen, dass die Ergebnisse unabhängig von den anderen Testergebnissen der Stellwerktests zu interpretieren sind. Mittelwert und Standardabweichung beziehen sich ausschliesslich auf die 5527 beteiligten Schülerinnen und Schüler der 8. Klassen im Schuljahr 2008 des Kantons St. Gallen.

Im Gegensatz zum adaptiven Testsystem Stellwerk, das die Ergebnisse ständig auf der gleichen, geeichten Skala (mit einheitlicher Metrik) ausweist, muss die Skala bei einem Schreibenanlass bei jeder Durchführung wieder neu berechnet beziehungsweise erstellt werden. Sowohl die Schwierigkeit der Texte als auch die Massstäbe der korrigierenden Lehrpersonen werden bei der nächsten Durchführung nicht exakt gleich sein.

Die durchschnittlichen Ergebnisse der Schülerinnen und Schüler der drei Schultypen (Kleinklasse, Realschule und Sekundarschule) entsprechen den Erwartungen. Allerdings sind die Unterschiede zwischen den Mittelwerten innerhalb der Sekundarschulen und innerhalb der Realschulen sehr gross. Viele Realschulklassen erreichen deutlich bessere Klassenmittelwerte als Sekundarschulklassen. Die Ergebnisse zeigen, dass eine schultypenunabhängige Beurteilung einer wertvollen Zusatzinformation entspricht, die sowohl für die Schülerinnen und Schüler als auch für die Lehrpersonen und die Eltern zu einer hilfreichen Information führt. Denn entscheidend für die Beurteilung ist einzig der Text, nicht aber der Schultyp oder die besuchte Klasse.



## Anhang 2

=====  
 Analyse Schreibanlass St.Gallen Fri Apr 18 13:24 2008  
 TABLES OF RESPONSE MODEL PARAMETER ESTIMATES  
 =====

TERM 1: rater

VARIABLES		UNWEIGHTED FIT			WEIGHTED FIT				
rater		ESTIMATE	ERROR <sup>^</sup>	MNSQ	CI	T	MNSQ	CI	T
1	1	-0.120	0.009	1.08 ( 0.94, 1.06)		2.6	1.06 ( 0.94, 1.06)		1.8
2	2	-0.178	0.011	1.17 ( 0.92, 1.08)		3.9	1.19 ( 0.92, 1.08)		4.1
3	3	0.272	0.010	0.89 ( 0.93, 1.07)		-3.3	0.89 ( 0.93, 1.07)		-3.2
4	4	0.027*	0.017	1.06 ( 0.92, 1.08)		1.4	1.08 ( 0.92, 1.08)		1.8

=====  
 An asterisk next to a parameter estimate indicates that it is constrained  
 Separation Reliability = 0.998  
 Chi-square test of parameter equality = 1221.94, df = 3, Sig Level = 0.000  
 ^ Quick standard errors have been used  
 =====

TERM 2: task

VARIABLES		UNWEIGHTED FIT			WEIGHTED FIT				
task		ESTIMATE	ERROR <sup>^</sup>	MNSQ	CI	T	MNSQ	CI	T
1	Handyverbot	0.083	0.010	1.04 ( 0.93, 1.07)		1.2	1.04 ( 0.93, 1.07)		1.0
2	Kampf den Kilos	-0.347	0.011	1.04 ( 0.91, 1.09)		0.8	1.03 ( 0.90, 1.10)		0.5
3	Schuluniform	0.228	0.011	1.02 ( 0.91, 1.09)		0.4	1.02 ( 0.89, 1.11)		0.4
4	Gefährl. Skifahren	0.217	0.012	1.02 ( 0.89, 1.11)		0.4	1.00 ( 0.87, 1.13)		-0.0
5	Anony.Bewerbung	-0.208	0.013	1.13 ( 0.87, 1.13)		1.8	1.15 ( 0.86, 1.14)		1.9
6	Spucken	0.026*	0.025	0.99 ( 0.92, 1.08)		-0.2	0.98 ( 0.91, 1.09)		-0.5

=====  
 An asterisk next to a parameter estimate indicates that it is constrained  
 Separation Reliability = 0.998  
 Chi-square test of parameter equality = 2075.44, df = 5, Sig Level = 0.000  
 ^ Quick standard errors have been used  
 =====