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 Pilotprojekts «Neugestaltung 3. Sek» und der Projektleitung der Bildungsdirektion des Kantons Zürich

Urs Moser

Zürich, 28. Februar 2010

Inhalt

1 Ausgangslage

Im Rahmen des Pilotversuchs «Neugestaltung des 9. Schuljahrs» wird das computergestützte Testsystem «Stellwerk» zur schultypenunabhängigen Leistungsbeurteilung eingesetzt. Die Stellwerk-Tests werden ausschliesslich am Computer gelöst. Es handelt sich um adaptive Tests, die sich den Fähigkeiten der Schülerinnen und Schüler anpassen. Adaptive Tests haben den Vorteil, 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 laufend ihre Fähigkeiten eingeschätzt. 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 fehlerlose 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 können deshalb am Computer nicht getestet werden. Aus diesem Grund hat die Projektleitung des Pilotversuchs «Neugestaltung des 9. Schuljahrs» das Institut für Bildungsevaluation der Universität Zürich mit der Durchführung eines Schreibanlasses zur Erfassung der sprachlichen Kompetenzen im Bereich «Texte schreiben» beauftragt.

2 Durchführung

Die Durchführung des Schreibanlasses fand am 14. und 15. Januar 2010 statt. Insgesamt verfassten 1126 Schülerinnen und Schüler der Pilotschulen einen Text. Zur Wahl standen zwei Themen: (1) «Macht Fernsehen dumm?» und (2) «Das Auto – Fluch oder Segen unserer Zeit?». Die Themen wurden vom Institut für Bildungsevaluation in Zusammenarbeit mit zwei Lehrpersonen entwickelt. Thema 1 wurde von 74 Prozent der Schülerinnen und Schüler gewählt, Thema 2 von 26 Prozent. Von den Knaben wählten 69 Prozent Thema 1 und 31 Prozent Thema 2, von den Mädchen wählten 81 Prozent Thema 1 und 19 Prozent Thema 2.

Die Themen wurden den Schülerinnen und Schülern mit kurzen Texten, die auf Zeitungsartikeln basieren, vorgestellt. Danach wurden zwei Aufgaben gestellt. Teilaufgabe 1 verlangte von den Schülerinnen und Schülern, die wichtigsten Aussagen des Textes in fünf Sätzen zusammenzufassen. Mit der zweiten Teilaufgabe wurde von den Schülerinnen und Schülern ein argumentativer Text mit drei Vorgaben verlangt: (1) Der Text musste mit einer kurzen Einleitung beginnen, die zum Thema hinführt. (2) Im Text mussten Argumente aufgeführt werden, die den Nutzen und die Gefahren des Fernsehens beziehungsweise des Autos aufzeigen. (3) Die Schülerinnen und Schüler mussten abschliessend Stellung nehmen, welche Einschätzung (Vor- oder Nachteile) für sie am ehesten zutrifft.

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

Bearbeite zu diesem Thema die zwei 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.*
- *Du darfst den Duden beziehungsweise das Wörterbuch benutzen.*

Die Texte der Schülerinnen und Schüler mussten von den Lehrpersonen kopiert und die Originale dem Institut für Bildungsevaluation zur Beurteilung zugestellt werden. Die Ergebnisse in Form einer Punktzahl wurden den Lehrpersonen anschliessend von der Firma «Arcadix», die im Projekt «Stellwerk» für die Informatik zuständig ist, auf dem Internet zur Verfügung gestellt. Ab Freitag, 10. Februar 2010, waren die Ergebnisse im Bereich «Texte schreiben» für die Lehrpersonen beziehungsweise für die Schülerinnen und Schüler als Teil des Stellwerk-Zertifikats auf dem Internet einsehbar.

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. Der Kriterienraster wurde in Anlehnung an den Zürcher Textanalyseraster von Nussbaumer & Sieber (1994)[1] 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)[2]. Die Beurteilung bezieht sich auf die kommunikativen und linguistischen Fähigkeiten. Zusätzlich wurden im Sinne einer ganzheitlichen Bewertung (Holistic Scoring)[3] die Textstruktur, der Sprachstil und das ästhetische Wagnis beziehungsweise die Kreativität beurteilt.

Tabelle 3.1 zeigt die Kriterien zur Beurteilung der kommunikativen Fähigkeiten. Aufgabe 1 verlangte die Zusammenfassung des Zeitungsartikels in fünf Sätzen. Beurteilt wurde, ob der Inhalt des Artikels korrekt und vollständig wiedergegeben wurde. Aufgabe 2 verlangte eine Einleitung, Argumente sowie eine abschliessende Stellungnahme.

---

[1] 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.

[2] Weigle, S. C. (2002). *Assessing Writing*. Cambridge. Cambridge University Press.

[3] Weigle, S. C. (2002). *Assessing Writing*. Cambridge. Cambridge University Press.

Tabelle 3.2 zeigt die Kriterien zur Beurteilung der linguistischen Fähigkeiten. Beurteilt wurden Wortwahl, Rechtschreibung und Zeichensetzung sowie Grammatik und Syntax.

Tabelle 3.3 zeigt drei Kriterien, die für eine Gesamtbeurteilung genutzt wurden. Beurteilt wurde, ob die Argumente in einen zusammenhängenden Text eingebettet wurden (Textstruktur), wie gewandt sich die Schülerinnen und Schüler sprachlich ausdrücken konnten (Stil) und ob sie im Text auch etwas wagten. Zudem wurde die Textlänge beurteilt.

Die Kriterien wurden jeweils anhand von zwei, drei, vier oder fünf Kategorien umschrieben. Das heisst, dass je nach Kriterium zwischen 0 und 4 Punkten vergeben werden konnten. Insgesamt konnten die Schülerinnen und Schüler 41 Punkte erreichen. Die einzelnen Kriterien wurden gleich wie Testaufgaben behandelt und einer Itemanalyse nach der klassischen Testtheorie unterzogen.

In der ersten Spalte der Tabellen 3.1, 3.2 und 3.3 befindet sich die Bezeichnung des Beurteilungskriteriums, in der zweiten Spalte die qualitative oder quantitative Abstufung zur Beurteilung des Textes anhand des Kriteriums. In der dritten Spalte befindet sich der Prozentanteil der Texte, denen die Ausprägung des Kriteriums zugeordnet wurde. In 6 Prozent der Texte war beispielsweise bei der ersten Aufgabe der Inhalt des Textes nicht wiedergegeben, in 41 Prozent der Texte war der Inhalt des Textes nur teilweise zusammengefasst und in 53 Prozent war der Inhalt vollständig wiedergegeben (siehe Tabelle 3.1).

Tabelle 3.1: Kriterien zur Beurteilung der kommunikativen Fähigkeiten

Beurteilungskriterium	Quantitative Abstufungen	Prozentanteil Texte	Trennschärfe
Aufgabe 1: Zusammenfassung			
– Wiedergabe des Inhaltes	nicht erfüllt	6%	0.29
	teilweise erfüllt	41%	
	vollständig erfüllt	53%	
– Hauptaussagen enthalten	keine Hauptaussage erwähnt	3%	0.35
	eine Hauptaussage erwähnt	5%	
	zwei bis drei Hauptaussagen erwähnt	34%	
	vier Hauptaussagen erwähnt	58%	
Aufgabe 2: Fragen beantworten			
– Argumente	nicht genannt	3%	0.54
	teilweise genannt	22%	
	genannt	75%	
– Stellungnahme	nicht vorhanden	10%	0.50
	teilweise vorhanden	20%	
	vollständig vorhanden	70%	
Aufgabe 2: Textform			
– Einleitung	nicht vorhanden	9%	0.63
	teilweise vorhanden	17%	
	vollständig vorhanden	74%	
– Ende	nicht vorhanden	21%	0.64
	teilweise vorhanden	38%	
	vollständig vorhanden	41%	

Tabelle 3.2: Kriterien zur Beurteilung der linguistischen Fähigkeiten

Beurteilungskriterium	Qualitative Abstufungen	Prozentanteil Texte	Trennschärfe
Aufgaben 1 und 2			
– Wortwahl	einfach, simpel	3%	0.42
	adäquat	92%	
	elaboriert, herausragend, überraschend	5%	
– Gross- und Kleinschreibung	kaum beherrscht	11%	0.50
	teilweise beherrscht	33%	
	nahezu fehlerfrei	56%	
– Rechtschreibung insgesamt	kaum beherrscht	11%	0.51
	teilweise beherrscht	41%	
	nahezu fehlerfrei	48%	
– Satzzeichen	rudimentär vorhanden (Punkt, kein Komma)	26%	0.61
	meist korrekte Satzzeichensetzung	46%	
	(nahezu) fehlerfrei	28%	
– Grammatik: Fallformen	kaum beherrscht	4%	0.48
	teilweise beherrscht	18%	
	nahezu fehlerfrei	78%	
– Syntax: Satzverbindung	keine, einfache Sätze	2%	0.66
	immer gleich	68%	
	abwechslungsreich	30%	
– Syntax: allgemein	teilweise korrekte Sätze	5%	0.76
	einfache korrekte Sätze	53%	
	komplexe korrekte Sätze (HS und NS)	42%	

Tabelle 3.3: Kriterien zur Gesamtbeurteilung

Beurteilungskriterium	Qualitative Abstufungen	Prozentanteil Texte	Trennschärfe
Aufgaben 1 und 2			
– Textstruktur	Textkerne (unverbunden)	6%	0.78
	eindimensionaler Text (logische Verkettung)	29%	
	mehrdimensionaler Text (gegliedert)	45%	
	mehrdimensionaler Text (abgeschlossen)	20%	
– Sprachstil	sprachlich unsicher	9%	0.79
	klar, aber einfache Sprachstrukturen	46%	
	sprachlich gewandt	43%	
	sprachlich sehr gewandt, ausdrucksstark	2%	
– Ästhetisches Wagnis/Kreativität	wagt wenig, einfache Lösung	6%	0.63
	wagt etwas, Kreativität erkennbar	29%	
	wagt viel, kreativ	45%	
	unerwartete Ideen, ausgesprochen kreativ	20%	
– Textlänge	weniger als eine halbe Seite	0%	0.66
	eine halbe Seite	1%	
	eine Seite	10%	
	eineinhalb Seiten	27%	
	mehr als eineinhalb Seiten	62%	

In der vierten Spalte der Tabellen 3.1, 3.2 und 3.3 sind die Angaben zur Trennschärfe des Kriteriums enthalten. Der Trennschärfekoeffizient zeigt bei einem Test, inwiefern eine Aufgabe Schülerinnen und Schüler mit hoher Punktzahl von Schülerinnen und Schülern mit niedriger Punktzahl 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 Itemanalyse zeigt, dass die Kriterien zur Beurteilung der linguistischen Fähigkeiten sowie jene zur Gesamtbeurteilung wesentlich besser zwischen guten und weniger guten Texten beziehungsweise Schülerinnen und Schülern differenzieren als die Kriterien zu den kommunikativen Fähigkeiten. Besonders gross ist die Differenzierung bei der Beurteilung der Textstruktur und des Sprachstils.

Am strengsten beurteilt wurden der Wortschatz, der Sprachstil, die Kreativität, die Syntax und die Textstruktur (beziehungsweise eher selten erreichten die Schülerinnen und Schüler bei diesen Kriterien die höchste Punktzahl). Am mildesten beurteilt beziehungsweise eher gut erfüllt wurden die Verständlichkeit sowie die inhaltliche Wiedergabe der Zusammenfassung.

Die Reliabilität beziehungsweise die Messgenauigkeit erreicht ein Cronbach-Alpha von $\alpha = 85$, was darauf hinweist, dass die Beurteilungskriterien ziemlich konsistent angewendet wurden und sich relativ gut eigneten, um zuverlässig zwischen guten und weniger guten Texten zu unterscheiden[4]. Wie zuverlässig die Korrektur der Texte durchgeführt wurde, ist im folgenden Kapitel 4 detailliert beschrieben.

4 Beurteilungszuverlässigkeit

Die Texte wurden von zwei erfahrenen Lehrpersonen mit Germanistikstudium (Rater) nach den vorgegebenen Kriterien korrigiert und beurteilt. Die beiden Rater arbeiten bereits seit mehreren Jahren am Institut für Bildungsevaluation bei der Korrektur von Texten mit und sind im Korrigieren von Texten nach dem vorgegebenen Kriterienkatalog versiert. Das Korrekturteam wurde bewusst klein gehalten, sodass die Standardisierung der Beurteilung dank gemeinsamer Absprache und regelmässiger Kontrolle hochgehalten werden konnte.

---

[4] Die Reliabilität beziehungsweise die Zuverlässigkeit eines Tests bezeichnet die Genauigkeit, mit der ein Test eine Personeneigenschaft misst. Die Reliabilität einer Textbeurteilung kann auch als die Replizierbarkeit der Textbeurteilung verstanden werden, die mittels eines Korrelationskoeffizienten angegeben wird. Bei einmaliger Testvorgabe wird zur Berechnung der Reliabilität der Koeffizient «Cronbach-Alpha» verwendet (Lienert, G. A. & Raatz, U. [1994]. *Testaufbau und Testanalyse*. Basel: Beltz.

In einer ersten Schulungsphase wurde der Kriterienkatalog auf seine Tauglichkeit überprüft. Anhand einer zufälligen 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 beiden beurteilenden Personen.

Während der gesamten Korrekturphase wurden 100 Texte doppelt korrigiert, um die Beurteilungsübereinstimmung ständig zu überprüfen. Abweichungen in der Beurteilung wurden laufend diskutiert mit dem Ziel, die Beurteilungsübereinstimmung (Inter-Rater-Reliabilität) hochzuhalten. Tabelle 4.1 enthält Informationen zur Beurteilungsübereinstimmung von den während der Korrekturphase 100 doppelt korrigierten Texten. In der zweiten Spalte ist pro Kriterium angegeben, wie hoch die Übereinstimmung in Prozent ist. Die Spalten 3 und 4 enthalten die Prozentanteile der Abweichungen nach Punkten.

Tabelle 4.1: Prozentuale Übereinstimmung und Inter-Rater-Reliabilität

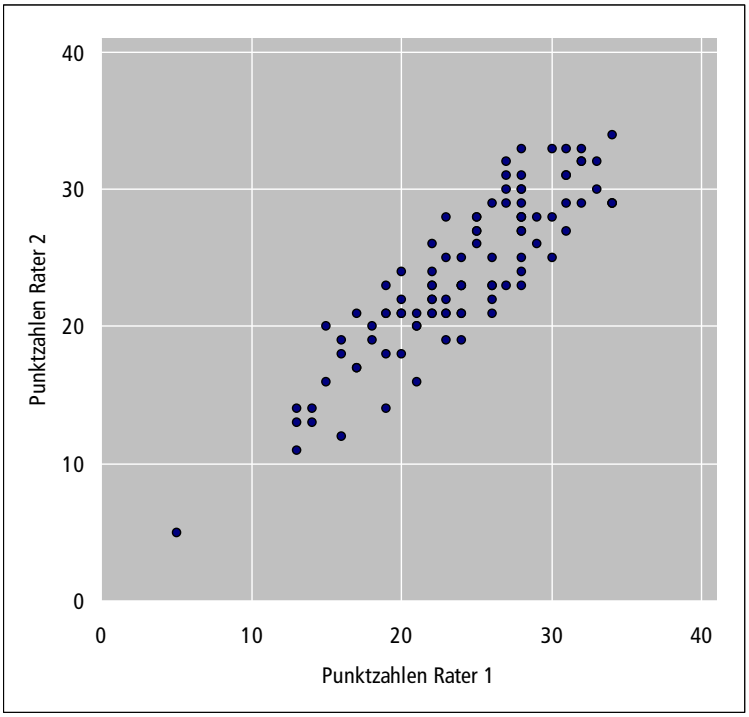
Beurteilungskriterium	Übereinstimmung	Abweichung: 1 Punkt	Abweichung: 2 Punkte	Kappa
Wiedergabe des Textes	87%	13%		0.64
Hauptaussagen enthalten	76%	24%		0.59
Argumente	64%	36%		0.51
Stellungnahme	69%	31%		0.55
Anfang	71%	29%		0.62
Ende	74%	26%		0.80
Textlänge	88%	12%		0.85
Wortwahl	87%	13%		0.56
Gross- und Kleinschreibung	70%	30%		0.63
Satzzeichen	59%	41%		0.40
Rechtschreibung insgesamt	67%	33%		0.57
Grammatik: Fallformen	64%	36%		0.32
Syntax: Satzverbindungen	87%	13%		0.83
Syntax: allgemein	69%	31%		0.59
Verständlichkeit	67%	33%		0.48
Textstruktur	71%	26%	3%	0.66
Sprachstil	67%	30%	3%	0.55
Ästhetisches Wagnis/Kreativität	72%	26%	2%	0.70

Die prozentuale Übereinstimmung ist – wie bereits beim Durchgang 2009 – bei der Beurteilung der Textlänge am höchsten und bei der Beurteilung der Satzzeichen am geringsten. Indem die Anzahl der Übereinstimmungen berechnet und am Anteil der zufälligen Übereinstimmung relativiert wird, kann das statistische Zusammenhangsmass «Kappa»

zur Bestimmung der Beurteilungsübereinstimmung berechnet werden (Tabelle 4.1)[5]. Der Kappa-Koeffizient kann Werte zwischen –1 und +1 annehmen. Der maximale Wert wird bei totaler Übereinstimmung erreicht. Bei einer systematisch gegensätzlichen Einstufung wird der Koeffizient negativ. Das Kappa hängt unter anderem auch von der Anzahl Abstufungen eines Kriteriums ab. Ist der Kappa-Koeffizient > 0.70, dann wird die Übereinstimmung als gut bezeichnet.

Der Kappa-Koeffizient wird bereits bei relativ hoher Übereinstimmung in der Beurteilung tief, was mit Vorsicht zu interpretieren ist. Um sich ein besseres Bild von der Beurteilungszuverlässigkeit machen zu können, wurden die Gesamtergebnisse der doppelt korrigierten Texte miteinander verglichen. Abbildung 4.1 zeigt die Punktzahlen der 100 doppelt korrigierten Texte. Die Position eines Punktes ergibt sich aus der Punktzahl von Rater 1 und der Punktzahl von Rater 2. Insgesamt konnten 41 Punkte vergeben werden. Die Korrelation zwischen den Beurteilungen der beiden Rater ist hoch und beträgt r = 0.89.

Abbildung 4.1: Differenzen in der Beurteilung der gleichen Texte durch zwei Rater



Bei 16 der 100 doppelt korrigierten Texte verteilten die Rater exakt gleich viele Punkte. Bei 26 Texten betrug die Abweichung 1 Punkt, bei 19 Texten 2 Punkte, bei 15 Texten 3 Punkte bei 11 Texten 4 Punkte und bei 13 Texten fünf Punkte.

---

[5] Zuerst wird der Anteil der beobachteten Übereinstimmungen P_0 berechnet (Diagonale in einer k-mal-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)].

Weil in einer Klasse fälschlicherweise alle Schülerinnen und Schüler zu beiden Themen einen Text verfassten, ergab sich die Möglichkeit, die Stabilität der Korrektur auch anhand von zwei Textbeispielen des gleichen Schülers beziehungsweise der gleichen Schülerin zu überprüfen. Mit einem Korrelationskoeffizienten von r = 0.90 war die Übereinstimmung insgesamt sogar noch leicht höher als jene der Doppelkorrekturen.

Das Beurteilungsverfahren kann als stabil und zuverlässig bezeichnet werden. Trotz zum Teil nicht optimaler Übereinstimmung bei einigen Beurteilungskriterien lassen sich mit dem gewählten Beurteilungsverfahren die sprachlichen Kompetenzen ebenso zuverlässig erfassen wie mit Tests, die sich auf gebundene Aufgabenformate wie Multiple-Choice-Aufgaben beschränken.

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 Folgen auf die Ergebnisse der Schülerinnen und Schüler haben?

Unterschiedliche Beurteilungsmassstäbe können mit verschieden 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 der gleiche Schreibanlass je nach beurteilender Person entweder etwas einfacher oder etwas schwieriger.

Bei der Beurteilung eines Textes bestimmen 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 bei der Anwendung des Kriteriums generell eher niedrige Beurteilungen erhalten. Dies trifft beispielsweise für das Kriterium «sprachlich sehr gewandt, ausdrucksstark» zu. (3) *Die Strenge oder Milde der beurteilenden Person (Rater).* Die Kriterien werden von den Ratern jeweils nicht exakt gleich interpretiert. (4) *Das Thema*. Texte zu spezifischen Themen werden nicht immer gleich streng beurteilt.

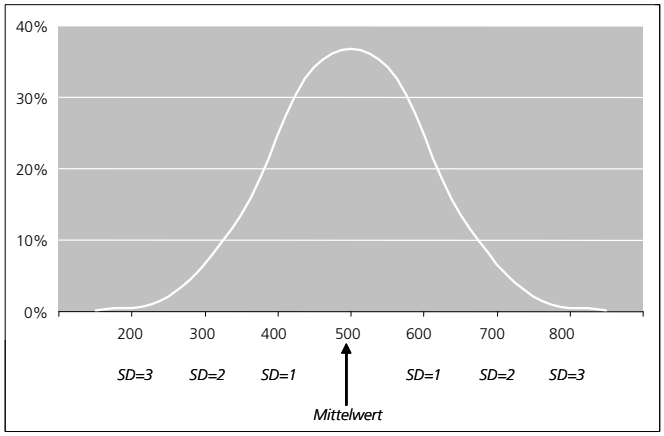
Die Fähigkeit der Schülerinnen und Schüler, die Beurteilungsstrenge der Rater und das Thema bestimmen das Ergebnis der Schülerinnen und Schüler. Sie werden deshalb als Facetten der Urteilssituation aufgefasst und bei der Berechnung der Ergebnisse berücksichtigt[6]. Mit der Anwendung der Item-Response-Theorie ist es möglich, die Beurteilungsstrenge der beurteilenden Personen sowie das Thema ins Testmodell einzubeziehen und

[6] 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.

bei der Berechnung der Ergebnisse entsprechend zu berücksichtigen. Ein solches Vorgehen wird auch als «Multi-Faceted Measurement» oder als «Multi-Facetten-Modell» bezeichnet[7]. 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[8].

Die Analyse zeigt, dass die beiden beurteilenden Personen (Rater) einen vergleichbaren Massstab anwenden (Anhang 1 und 2). Die beiden Zahlen «1» und «2» in der Spalte «+rater» im Anhang 1 liegen nahe beieinander. Die Tabelle im Anhang 2 zeigt, dass die Differenz bei der Beurteilung der gleichen Texte bei rund ± 0.063 Logits liegt, was bei der Berechnung der Ergebnisse der Schülerinnen und Schüler berücksichtigt wird. Auch die Wahl des Themas ist eher von untergeordneter Bedeutung (± 0.021). Das Thema «Macht Fernsehen dumm?» wurde etwas milder beurteilt als das Thema «Das Auto – Fluch oder Segen unserer Zeit?» (Anhang 2).

Abbildung 5.1: Verteilung der Testergebnisse



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 mussten. Dabei wurden die Testrohwerte so transformiert, dass – analog der Stellwerk-Skala – der Mittelwert 500 Punkte und die Standardabweichung 100 Punkte betragen (vgl. Abbildung 5.1). 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. Die Anzahl Punkte zeigt den Schülerinnen und Schülern, wie gut sie innerhalb der Vergleichsgruppe – 1126 Schülerinnen und Schüler, die den Text geschrieben haben – abgeschnitten haben.

---

[7] McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

[8] Rost, J. (2003). *Lehrbuch Testtheorie – Testkonstruktion.* Bern: Hans Huber.

6 Kompetenzen und Textbeispiele

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[9]. Tabelle 6.1 zeigt zusammenfassend, welche Schreibkompetenzen innerhalb eines bestimmten Intervalls vorhanden sind.

Tabelle 6.1: Kompetenzbeschreibungen nach Punkteintervallen

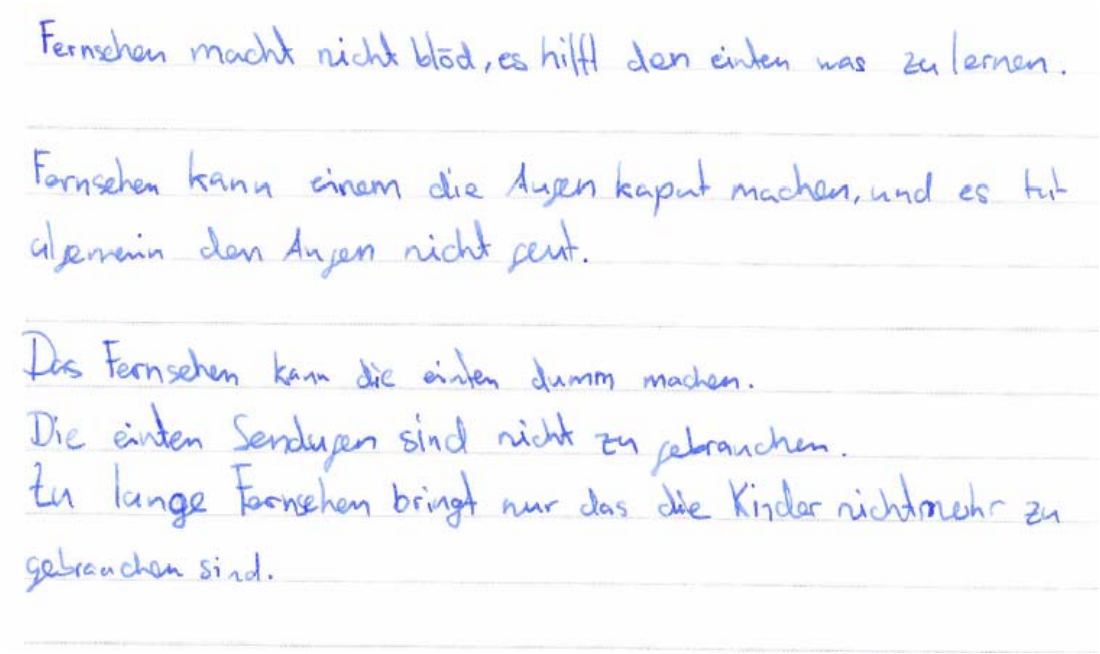
Punkteintervall	Kompetenzbeschreibungen
200 bis 300 Punkte	Die Texte sind eher kurz (eine Seite), aber grösstenteils verständlich. Die Zusammenfassung enthält eine Hauptaussage und gibt den Inhalt teilweise korrekt wieder. Die Rechtschreibung wird nicht beachtet. Die sprachliche Ausdrucksweise wird als unsicher beurteilt und die Texte bestehen aus einfachen Textkernen, die assoziativ miteinander verbunden sind. Argumente und Stellungnahmen sind erkennbar.
301 bis 400 Punkte	Die Texte sind jeweils eineinhalb Seiten lang. Die Zusammenfassung enthält zwei bis drei Hauptaussagen, jedoch ohne vollständige Wiedergabe des Inhalts. Anfang und Ende des Textes sind teilweise als solche erkennbar. Die Rechtschreibregeln werden zum Teil beachtet. Der Sprachstil ist einfach und klar. Argumente und Stellungnahmen sind vorhanden.
401 bis 500 Punkte	Die Texte sind mehr als eineinhalb Seiten lang. Sie sind gut verständlich und enthalten einfache, korrekte Sätze. Anfang und Ende der Texte sind vorhanden und die Textelemente werden sachlogisch miteinander verbunden, sodass eine Textgliederung erkennbar ist. Die Rechtschreibung wird teilweise beherrscht. Satzzeichen und Fallformen werden meist korrekt angewendet.
501 bis 600 Punkte	Die Texte enthalten vergleichsweise komplexe Sätze, die abwechslungsreich miteinander verbunden sind. Die Rechtschreibung ist nahezu fehlerfrei. Der Sprachstil ist gewandt. Die kommunikativen Aufgaben werden vollständig gelöst.
601 bis 700 Punkte	Die Texte sind sowohl in Bezug auf die Rechtschreibung als auch in Bezug auf die Satzzeichen nahezu fehlerfrei. Die Texte enthalten komplexe korrekte Sätze, sind sprachlich gewandt und kreativ. Auch die formalen Kriterien werden erfüllt. Anfang und Ende des Textes sind klar und der Text ist gegliedert.
701 bis 800 Punkte	Die Texte haben einen klaren Aufbau und sind in sich abgeschlossen. Die Texte sind ausgesprochen kreativ und überzeugen durch eine überraschende Wortwahl. Sie sind ausdrucksstark und werden als sprachlich gewandt beurteilt.

Die Intervalle sind hierarchisch aufgebaut. Das bedeutet für die Interpretation der Ergebnisse, dass Schülerinnen und Schüler, die ein 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 wurde, dann gilt für diesen Text selbstverständlich auch, dass die Satzzeichen meist korrekt angewendet werden.

[9] Moser, U. (2006). *Wie werden die Ergebnisse in den Stellwerk-Tests interpretiert? Von den Testergebnissen zu einer professionellen Beurteilung der Kompetenzen der Schülerinnen und Schüler.* (www.stellwerk-check.ch)

Textbeispiel 1 zeigt den Text eines Schülers, der mit 200 Punkten beurteilt wurde. Der Text ist grösstenteils verständlich und enthält im Ansatz Argumente «Zu lange Fernsehen bringt nur das die Kinder nicht mehr zu gebrauchen sind», wie das bei der Teilaufgabe 2 gefordert wird. Eine eigentliche Stellungnahme fehlt hingegen, auch wenn beispielsweise der Satz «Fernsehen macht nicht blöd, es hilft den einten was zu lernen» als solche interpretiert werden kann.

Textbeispiel 1: 200 Punkte: «Macht Fernsehen dumm?» (Teilaufgabe 2, ganzer Text)



Fernsehen macht nicht blöd, es hilft den einten was zu lernen.

Fernsehen kann einem die Augen kaput machen, und es tut allgemein den Augen nicht gut.

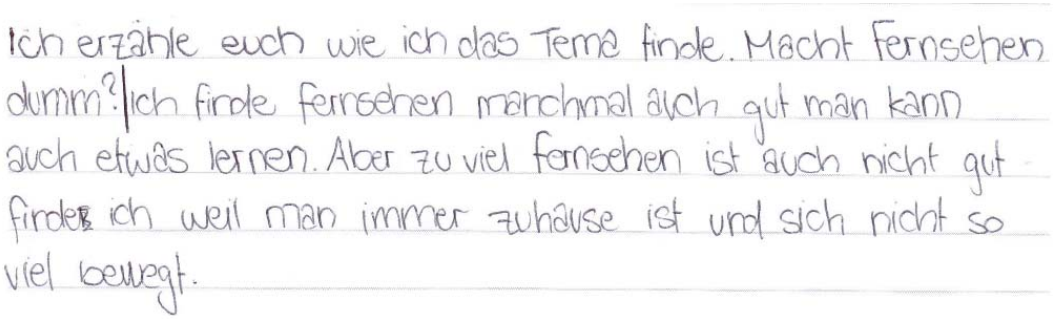
Das Fernsehen kann die einten dumm machen.
Die einten Sendungen sind nicht zu gebrauchen.
Zu lange Fernsehen bringt nur das die Kinder nichtmehr zu gebrauchen sind.

Die einzelnen Sätze sind korrekt, aber nicht miteinander verbunden. Der Text beginnt ohne Einleitung und er hat auch kein Ende. Mit einer knappen halben Seite ist der Text zudem sehr kurz. Die Rechtschreibung wird zwar teilweise beherrscht und auch Kommas werden richtig gesetzt. Allerdings lässt sich die Rechtschreibung aufgrund der wenigen Sätze und der einfachen Wortwahl nicht zuverlässig beurteilen. Der Text lässt auf eine sprachlich unsichere Ausdrucksweise schliessen.

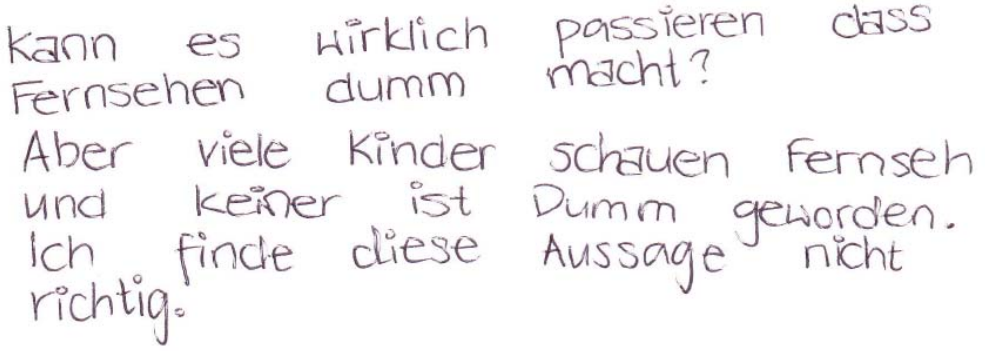
Die Textbeispiele 2 und 3 beziehen sich ebenfalls auf die zweite Teilaufgabe zum Thema «Macht Fernsehen dumm?». Die Texte sind leicht kürzer als das Textbeispiel 1. Trotzdem wurden die beiden Texte leicht besser beurteilt. Im Vergleich zu Textbeispiel 1 enthalten die beiden Textbeispiele ein Argument und eine Stellungnahme. Der kommunikative Auftrag ist rudimentär eingelöst. Beispielsweise wird das Fernsehen im Textbeispiel 2 positiv beurteilt, weil man auch etwas lernen kann («Ich finde fernsehen manchmal auch gut man kann auch etwas lernen»). Danach wird ein Argument aufgeführt, weshalb zu viel Fernsehen ungesund ist («Aber zu viel fernsehen ist auch nicht gut finde ich weil man immer zuhause ist und sich nicht so viel bewegt»).

Darüber hinaus sind erste Anzeichen einer sachlogischen Verkettung der Sätze vorhanden. Die Rechtschreibung wird teilweise beherrscht, obwohl auch dieses Kriterium bei der geringen Anzahl von Sätzen nicht zuverlässig beurteilt werden konnte. Die Satzzeichen sind nur rudimentär in Form von Satzschlusszeichen vorhanden.

Textbeispiel 2: 373 Punkte: «Macht Fernsehen dumm?» (Teilaufgabe 2, ganzer Text)

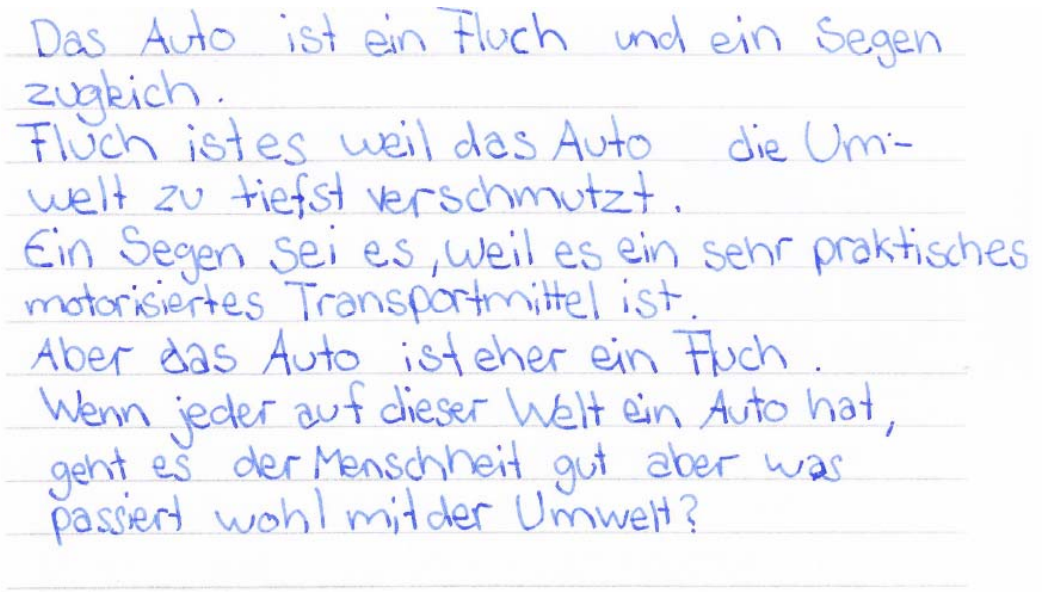

Ich erzähle euch wie ich das Teme finde. Macht Fernsehen dumm? Ich finde fernsehen manchmal auch gut man kann auch etwas lernen. Aber zu viel fernsehen ist auch nicht gut finde ich weil man immer zuhause ist und sich nicht so viel bewegt.

Textbeispiel 3: 373 Punkte: «Macht Fernsehen dumm?» (Teilaufgabe 2, ganzer Text)


kann es Hirklich passieren dass Fernsehen dumm macht?
Aber viele Kinder schauen Fernseh und keiner ist Dumm geworden. Ich finde diese Aussage nicht richtig.

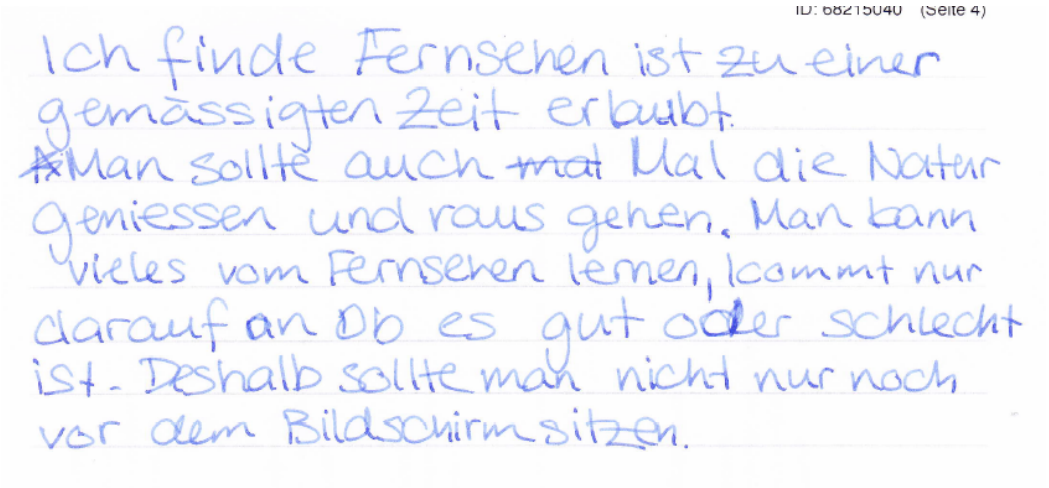
In den Textbeispielen 4 und 5 wird der kommunikative Auftrag wesentlich differenzierter erfüllt. In Textbeispiel 4 wird zuerst ein Argument gegen das Autofahren und dann ein Argument für das Autofahren aufgeführt. Im Anschluss folgt die Stellungnahme, die mit der Frage «Was wohl mit der Umwelt geschieht?» endet. Beim Textbeispiel 5 (Ausschnitt) handelt es sich um die Stellungnahme, die auf die Argumentation folgt. Beide Texte zeugen zudem von einer adäquaten Wortwahl. Die Rechtschreibung wird teilweise beherrscht. Die Satzzeichen sind meist korrekt gesetzt. Insgesamt handelt es sich um einfache sprachliche Lösungen, die zu einem in sich abgeschlossenen Text führen.

Textbeispiel 4: 479 Punkte: «Das Auto – Fluch oder Segen unserer Zeit?» (ganzer Text)



Das Auto ist ein Fluch und ein Segen zugleich.
Fluch ist es weil das Auto die Umwelt zu tiefst verschmutzt.
Ein Segen sei es, weil es ein sehr praktisches motorisiertes Transportmittel ist.
Aber das Auto ist eher ein Fluch.
Wenn jeder auf dieser Welt ein Auto hat, geht es der Menschheit gut aber was passiert wohl mit der Umwelt?

Textbeispiel 5: 479 Punkte: «Macht Fernsehen dumm?» (Teilaufgabe 2, Ausschnitt)

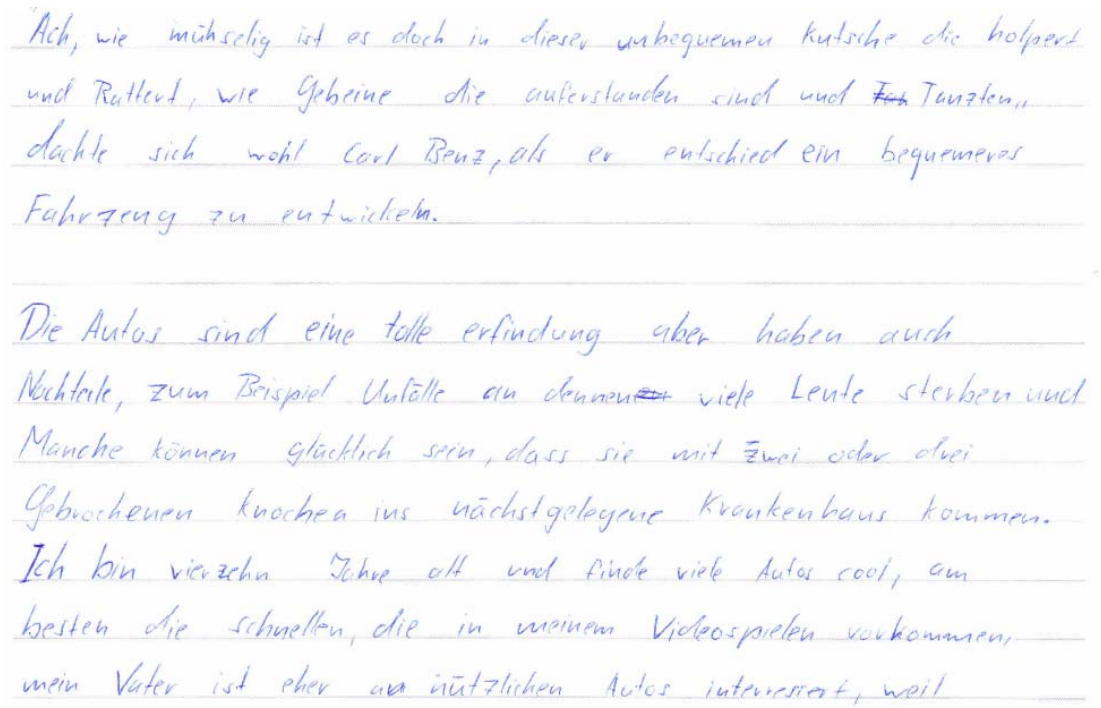


ID: 68215040 (Seite 4)

Ich finde Fernsehen ist zu einer gemässigten Zeit erlaubt.
Man sollte auch mal die Natur geniessen und raus gehen. Man kann vieles vom Fernsehen lernen, kommt nur darauf an ob es gut oder schlecht ist. Deshalb sollte man nicht nur noch vor dem Bildschirm sitzen.

Die Textbeispiele 6 und 7 unterscheiden sich von den Textbeispielen 5 und 6 vor allem im Aufbau und in der Komplexität. Die Texte enthalten einen Anfang. Textbeispiel 5 bezieht sich auf den Inhalt des Zeitungsartikels und schreibt eine originelle Einleitung («Ach, wie mühselig ist es doch in dieser unbequemen Kutsche die holpert und Rattert, …»). Textbeispiel 6 beginnt mit einem Einleitungssatz («Es gibt sehr viele Leute die über das Thema diskutieren ob Fernseher schauen dumm macht oder eben nicht»). Danach werden in beiden Texten verschiedene Argumenten aufgeführt, sodass eine gegliederte, mehrdimensionale Textstruktur erkennbar wird. Die Rechtschreibung wird teilweise beherrscht und auch die Satzzeichen werden unvollständig gesetzt. Im Vergleich zu den Textbeispielen 5 und 6 sind die Sätze allerdings komplexer, weshalb Rechtschreibung und Zeichensetzung im Verhältnis zur Komplexität des Textes beurteilt werden müssen. Die Schreibenden drücken sich sprachlich gewandt aus und sind dabei kreativ.

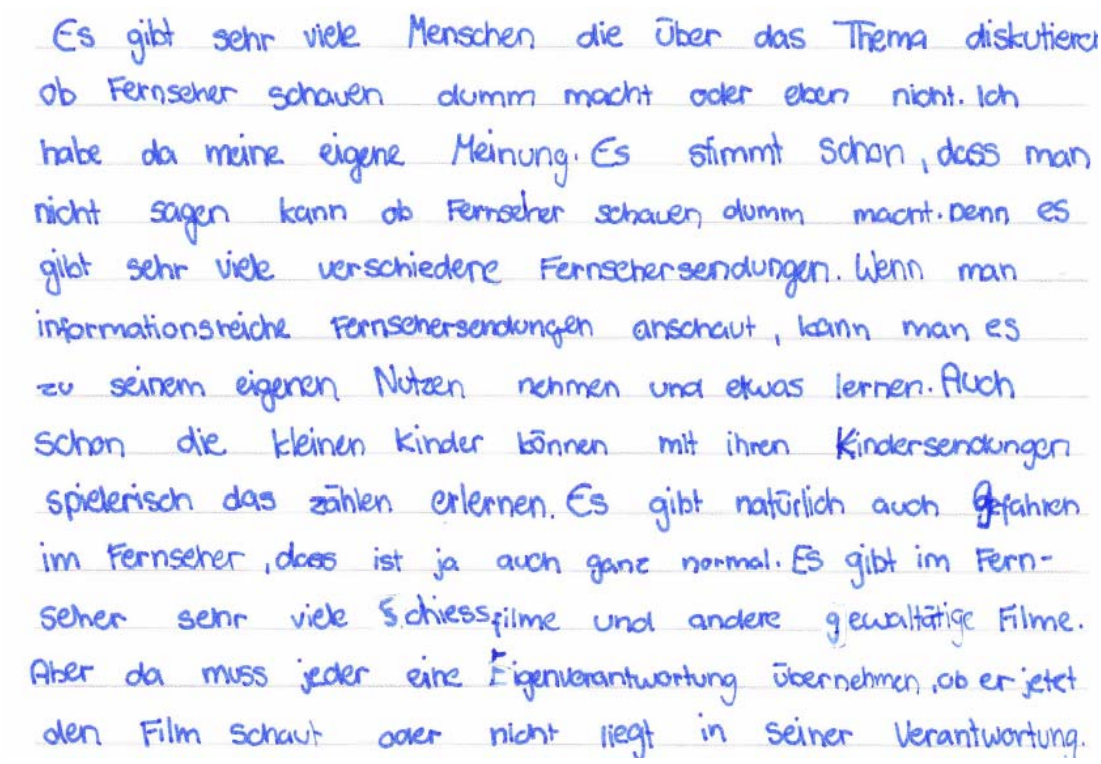
Textbeispiel 6: 523 Punkte: «Das Auto – Fluch oder Segen unserer Zeit?» (Ausschnitt)



Ach, wie mühselig ist es doch in dieser unbequemen Kutsche die holpert und Rattert, wie Gebeine die auferstanden sind und ~~Fuh~~ Tanzten, dachte sich wohl Carl Benz, als er entschied ein bequemeres Fahrzeug zu entwickeln.

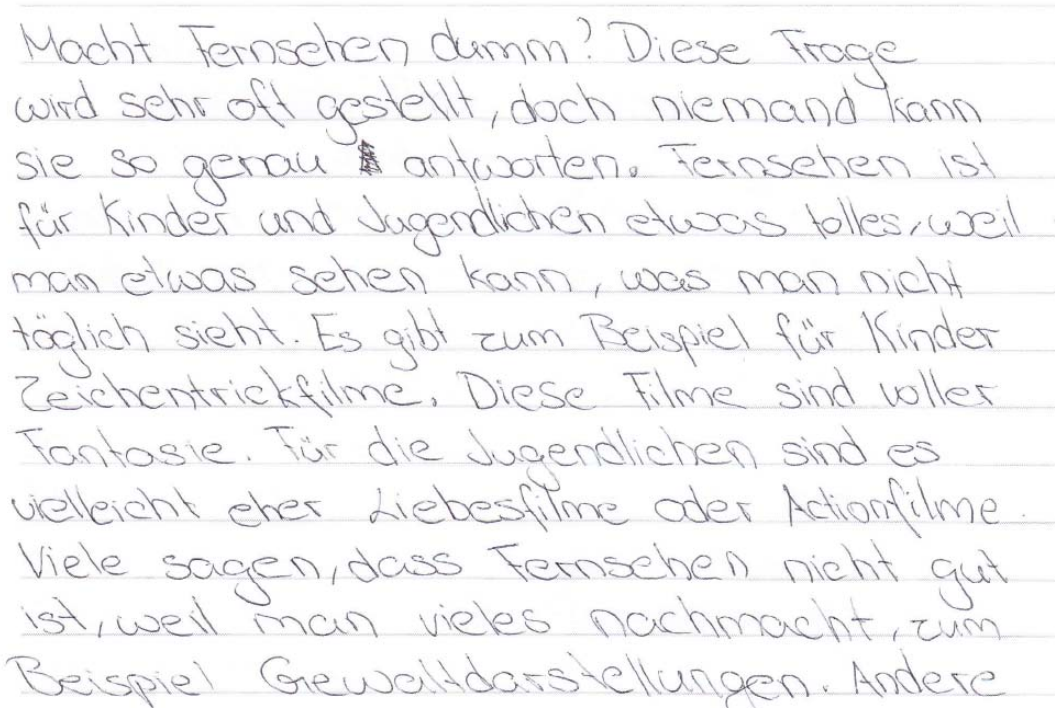
Die Autos sind eine tolle erfindung aber haben auch Nachteile, zum Beispiel Unfälle an denen viele Leute sterben und Manche können glücklich sein, dass sie mit Zwei oder drei Gebrochenen knochen ins nächstgelegene Krankenhaus kommen. Ich bin vierzehn Jahre alt und finde viele Autos cool, am besten die schnellen, die in meinem Videospielen vorkommen, mein Vater ist eher an nützlichen Autos interessiert, weil

Textbeispiel 7: 573 Punkte: «Macht Fernsehen dumm?» (Teilaufgabe 2, Ausschnitt)



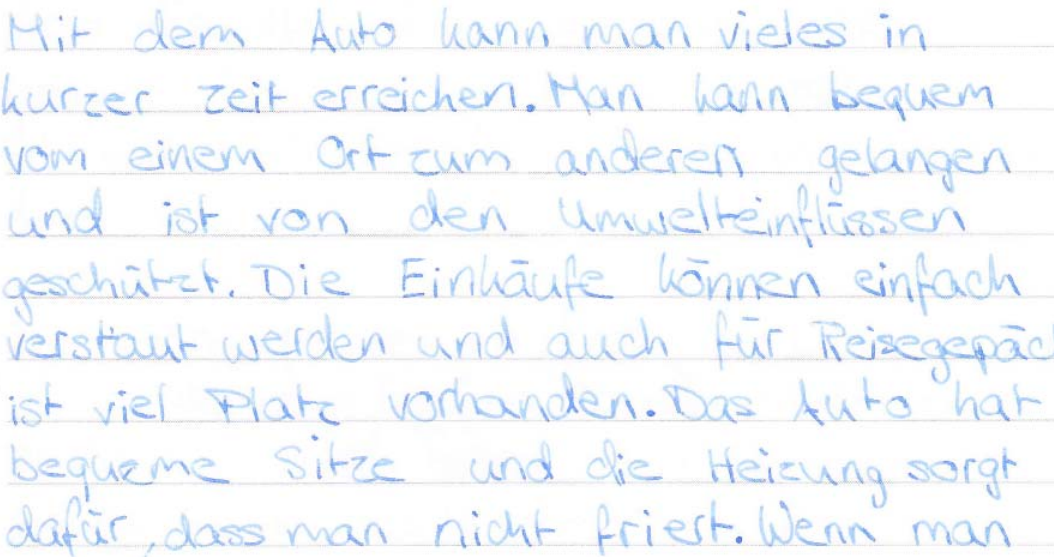
Es gibt sehr viele Menschen die über das Thema diskutieren ob Fernseher schauen dumm macht oder eben nicht. Ich habe da meine eigene Meinung. Es stimmt schon, dass man nicht sagen kann ob Fernseher schauen dumm macht. Denn es gibt sehr viele verschiedene Fernsehersendungen. Wenn man informationsreiche Fernsehersendungen anschaut, kann man es zu seinem eigenen Nutzen nehmen und etwas lernen. Auch schon die kleinen Kinder können mit ihren Kindersendungen spielerisch das zählen erlernen. Es gibt natürlich auch Gefahren im Fernseher, dass ist ja auch ganz normal. Es gibt im Fernseher sehr viele Schiessfilme und andere gewalttätige Filme. Aber da muss jeder eine Eigenverantwortung übernehmen, ob er jetzt den Film schaut oder nicht liegt in seiner Verantwortung.

Textbeispiel 8: 665 Punkte: «Das Auto – Fluch oder Segen unserer Zeit?» (Ausschnitt)



Macht Fernsehen dumm? Diese Frage wird sehr oft gestellt, doch niemand kann sie so genau antworten. Fernsehen ist für Kinder und Jugendlichen etwas tolles, weil man etwas sehen kann, was man nicht täglich sieht. Es gibt zum Beispiel für Kinder Zeichentrickfilme. Diese Filme sind voller Fantasie. Für die Jugendlichen sind es vielleicht eher Liebesfilme oder Actionfilme. Viele sagen, dass Fernsehen nicht gut ist, weil man vieles nachmacht, zum Beispiel Gewaltdarstellungen. Andere

Textbeispiel 9: 668 Punkte: «Das Auto – Fluch oder Segen unserer Zeit?» (Ausschnitt)

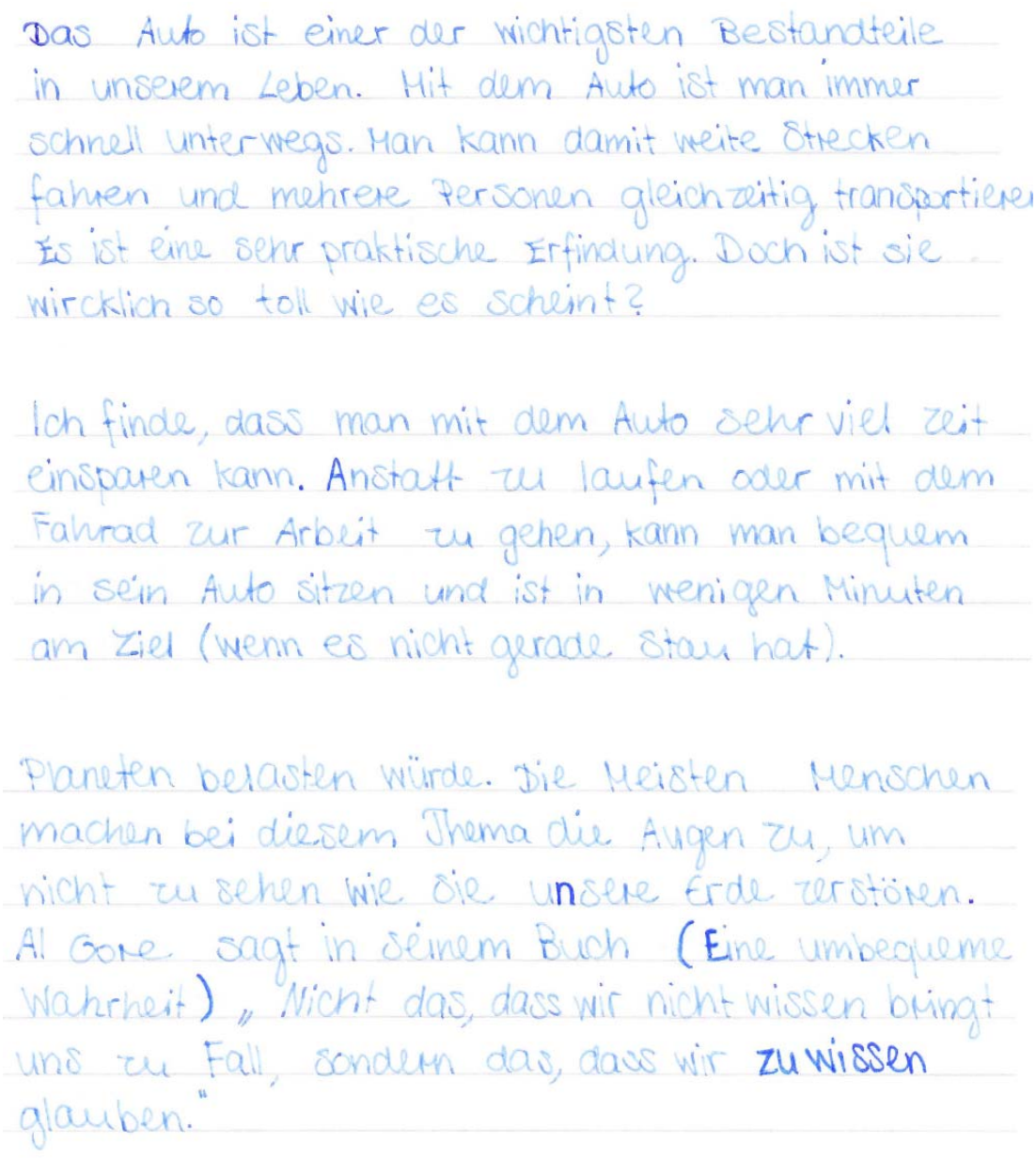


Mit dem Auto kann man vieles in kurzer Zeit erreichen. Man kann bequem vom einem Ort zum anderen gelangen und ist von den Umwelteinflüssen geschützt. Die Einkäufe können einfach verstaut werden und auch für Reisegepäck ist viel Platz vorhanden. Das Auto hat bequeme Sitze und die Heizung sorgt dafür, dass man nicht friert. Wenn man

Die Textbeispiele 8 und 9 unterscheiden sich von den Textbeispielen 6 und 7 vor allem in der Komplexität der Sätze und in der Wortwahl: «Fernsehen ist für Kinder und Jugendliche etwas tolles, weil man etwas sehen kann, was man nicht täglich sieht» oder «Man kann bequem von einem Ort zum anderen gelangen und ist von den Umwelteinflüssen geschützt». Die Texte sind in Bezug auf die Rechtschreibung und die Satzzeichen zwar nicht

vollständig, aber nahezu fehlerfrei. Weil nur Ausschnitte abgebildet sind, kommen andere Qualitäten der Texte hier nicht zum Ausdruck (beispielsweise Textstruktur, Anfang und Ende des Textes).

Textbeispiel 10: 800 Punkte: «Das Auto – Fluch oder Segen unserer Zeit?» (Ausschnitt)



Das Auto ist einer der wichtigsten Bestandteile in unserem Leben. Mit dem Auto ist man immer schnell unterwegs. Man kann damit weite Strecken fahren und mehrere Personen gleichzeitig transportieren. Es ist eine sehr praktische Erfindung. Doch ist sie wirklich so toll wie es scheint?

Ich finde, dass man mit dem Auto sehr viel Zeit einsparen kann. Anstatt zu laufen oder mit dem Fahrrad zur Arbeit zu gehen, kann man bequem in sein Auto sitzen und ist in wenigen Minuten am Ziel (wenn es nicht gerade Stau hat).

Planeten belasten würde. Die Meisten Menschen machen bei diesem Thema die Augen zu, um nicht zu sehen wie sie unsere Erde zerstören. Al Gore sagt in seinem Buch (Eine unbequeme Wahrheit) „Nicht das, dass wir nicht wissen bringt uns zu Fall, sondern das, dass wir zu wissen glauben."

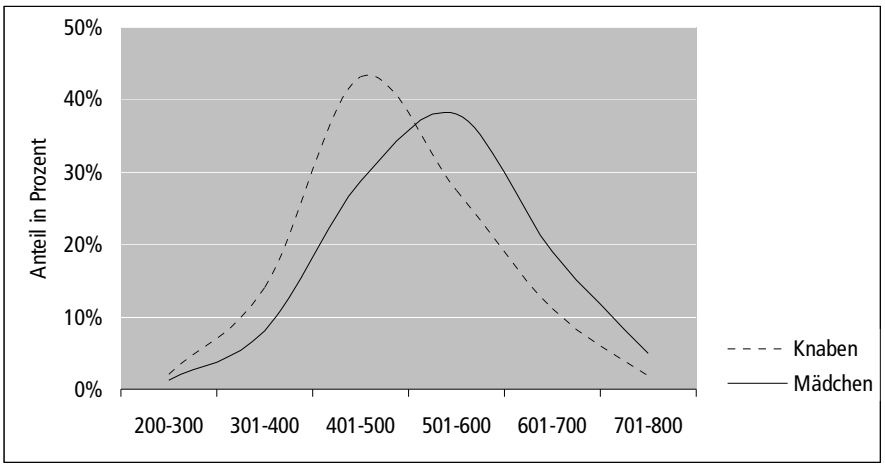
Textbeispiel 10 zeigt den Anfang, eine explizite Stellungnahme sowie den Schluss des Textes. Sprachlich gewandt wird ins Thema eingestiegen. Die Einleitung endet mit einer Frage, die Spannung auslöst. Die Stellungnahme ist klar und überzeugend. Der Text hat ein klares Ende. Die Textbeispiele überzeugen nicht nur durch den engagierten und ausgesprochen kreativen Zugang zum Thema. Trotz einiger Rechtschreibfehler ist der Text ausdrucksstark und lässt auf einen sprachlich gewandten Schüler schliessen.

7 Ergebnisse

7.1 Ergebnisse nach Geschlecht

Abbildung 7.1 zeigt die Verteilung der Ergebnisse nach Geschlecht. Der Mittelwert der Mädchen liegt bei 528 Punkten, der Mittelwert der Knaben bei 484 Punkten. Die Differenz von 44 Punkten – exakt gleich gross wie im letzten Jahr – ist statistisch signifikant und von mittlerer Bedeutung. Die Geschlechterdifferenzen sind nicht in allen Schultypen gleich gross. In der Abteilung A beträgt der Vorsprung der Mädchen 41 Punkte, in der Abteilung B beträgt der Vorsprung 27 Punkte und in der Abteilung C 31 Punkte.

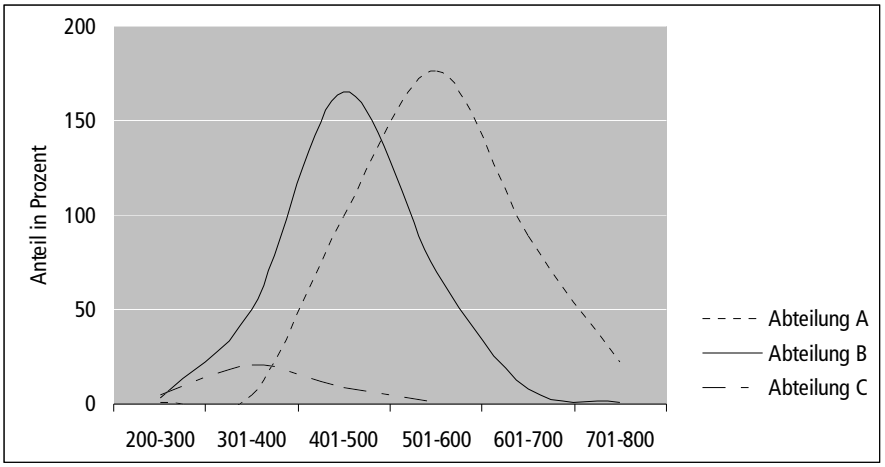
Abbildung 7.1: Ergebnisse nach Geschlecht



7.2 Ergebnisse nach Abteilungen

Abbildung 7.2 zeigt die Ergebnisse nach den Abteilungen (Schultypen). Die Verteilungskurven wurden aufgrund der Anzahl Schülerinnen und Schüler gebildet. Die Verteilungskurven entsprechen den Erwartungen.

Abbildung 7.2: Dreiteilige Sekundarschule nach Abteilungen

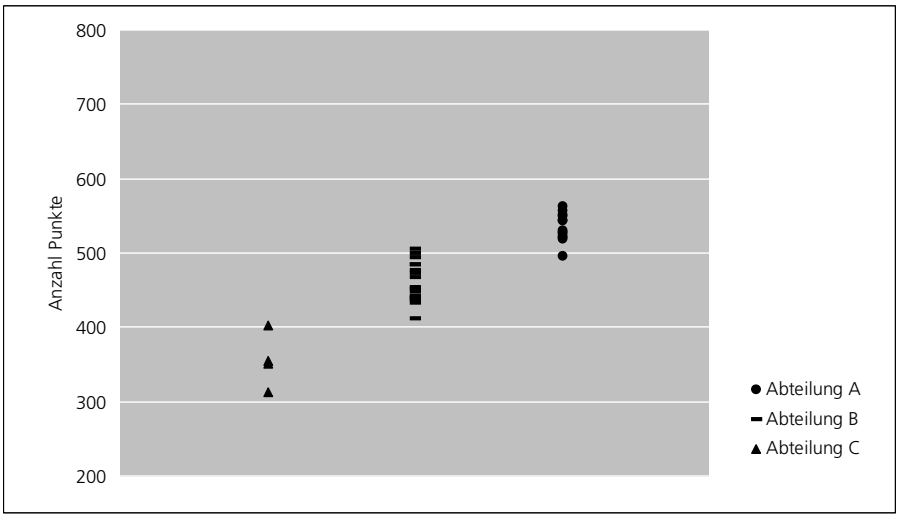


Die Texte der Schülerinnen und Schüler der Abteilung A wurden am häufigsten mit 501 bis 600 Punkten beurteilt (Mittelwert = 550 Punkte). Die Texte der Schülerinnen und Schüler der Abteilungen B wurden am häufigsten mit 401 bis 500 Punkten beurteilt (Mittelwert = 458 Punkte). Die Texte der Schülerinnen und Schüler der Abteilung C wurden am häufigsten mit 301 bis 400 Punkten beurteilt (Mittelwert = 368 Punkte). 6 Prozent der Schülerinnen und Schüler der Abteilung C schreiben Texte, deren Beurteilung über dem Mittelwert der Abteilung B liegt, rund 10 Prozent der Schülerinnen und Schüler der Abteilung B schreiben Texte, die über dem Mittelwert der Abteilung A liegen.

7.3 Ergebnisse nach Klassen

Abbildung 7.3 zeigt die Ergebnisse der beteiligten Klassen nach Abteilung (A, B und C). Die individuellen Testergebnisse wurden zu einem Klassenmittelwert zusammengefasst und sind in der Abbildung als Quadrat, Punkt oder Dreieck dargestellt. Die Klassenmittelwerte der Abteilungen C liegen zwischen rund 312 und 402 Punkten, die Mittelwerte der Abteilungen B liegen zwischen rund 411 und 505 Punkten und die Mittelwerte der Abteilungen A liegen zwischen rund 495 und 595 Punkten. Die Verteilung der Klassenmittelwerte zeigt das erwartete Bild. Die Klassen der Abteilung A erreichen in der Regel bessere Ergebnisse als jene der Abteilung B und die Klassen der Abteilung B erreichen in der Regel bessere Ergebnisse als jene der Abteilung C.

Abbildung 7.3: Klassenmittelwerte nach Abteilung



8 Fazit

Im Januar 2010 führten die Klassen des Pilotversuchs «Neugestaltung des 9. Schuljahrs» einen Schreibanlass durch, der extern mit einem standardisierten Verfahren korrigiert wurde. Die Durchführung verlief ohne Probleme.

Die Schülerinnen und Schülern konnten zwischen zwei Themen wählen: (1) «Macht Fernsehen dumm?» und (2) «Das Auto – Fluch oder Segen unserer Zeit?». Zu den Themen wurden zwei Aufgaben gestellt. Zum einen musste ein einleitender Zeitungsartikel mit fünf Sätzen zusammengefasst werden. Zum andern wurde ein argumentativer Text verlangt, der eine Einleitung hatte, Argumente enthielt und mit einer Stellungnahme abgeschlossen wurde.

Die Beurteilungsübereinstimmung der beiden korrigierenden Lehrpersonen (Rater) wurde mit verschiedenen Verfahren überprüft. Es zeigte sich, dass mit dem standardisierten Beurteilungsverfahren sehr zuverlässig beurteilt werden kann und die Resultate aus einer testtheoretischen Perspektive als sehr zuverlässig bezeichnet werden können.

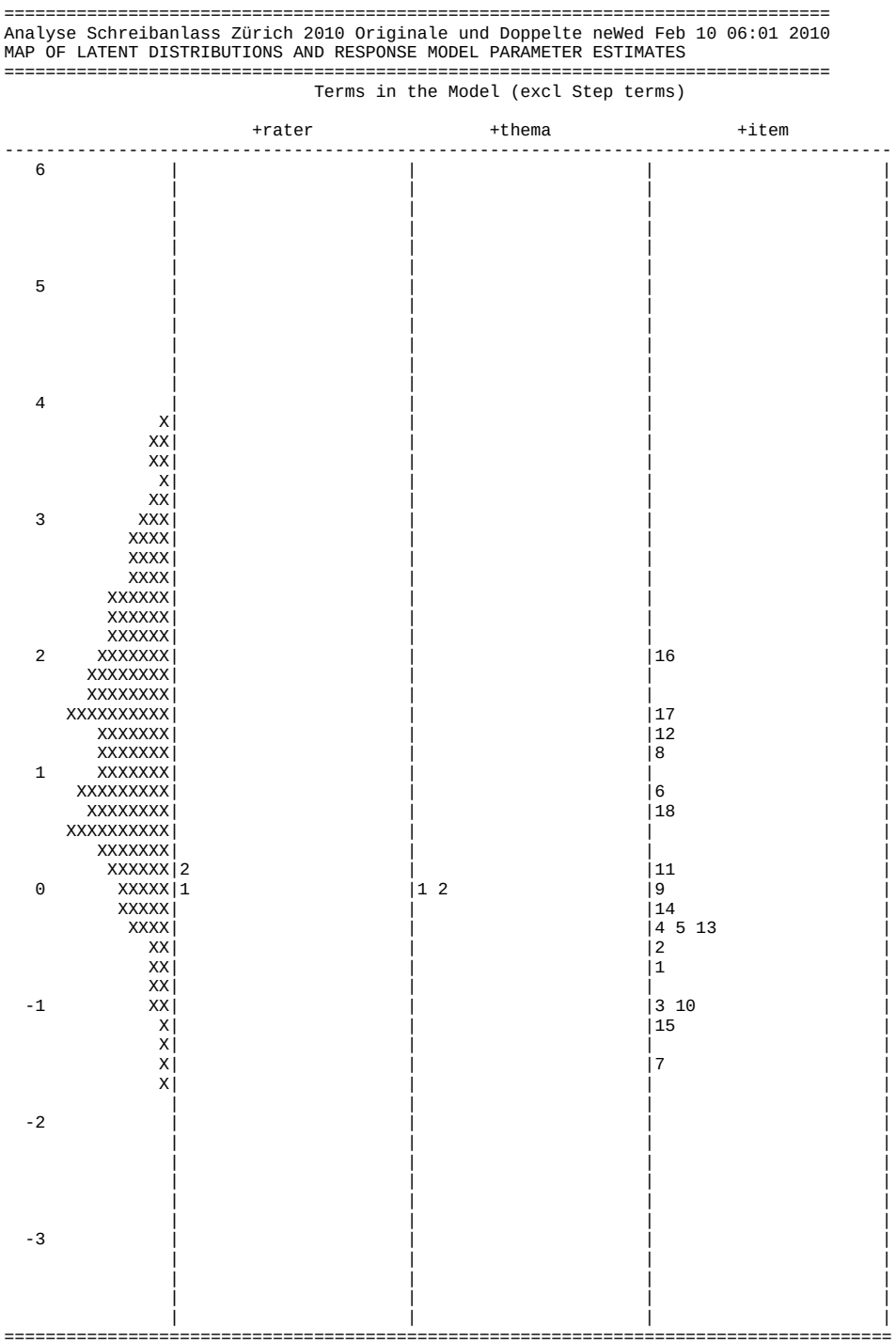
Dank der Anwendung der Item-Response-Theorie konnten die leicht unterschiedlichen Beurteilungsmassstäbe der Rater bei der Berechnung der Ergebnisse der Schülerinnen und Schüler korrigiert werden, sodass eine faire Beurteilung möglich wurde und einzelne Schülerinnen und Schüler nicht etwa aufgrund der korrigierenden Person oder des gewählten Themas benachteiligt waren.

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 Stellwerk-Tests zu interpretieren sind. Mittelwert und Standardabweichung beziehen sich ausschliesslich auf die 1126 beteiligten Schülerinnen und Schüler des Pilotversuchs des Kantons Zürich.

Textbeispiele der Schülerinnen und Schüler zeigen, dass das Spektrum in der schriftlichen Ausdrucksfähigkeit gross ist. Für rund 16 Prozent der Schülerinnen und Schüler stellte die Schreibaufgabe eine Überforderung dar: Selbst die kommunikativen Aufgaben wie das korrekte Zusammenfassen oder das Argumentieren wurden nur teilweise erfüllt. Mädchen erreichen im Durchschnitt statistisch signifikant und deutlich bessere Ergebnisse als Knaben. Die Darstellung der Ergebnisse nach den Abteilungen zeigt zudem, wie sinnvoll eine schultypenunabhängige Beurteilung sein kann. Rund 10 Prozent der Texte von Schülerinnen und Schülern der Abteilung B wurden beispielsweise höher bewertet als ein durchschnittlicher Text der Abteilung A.

Die Erfassung produktiver Kompetenzen ist aus einer testtheoretischen Perspektive gleich zuverlässig möglich wie die Erfassung reproduktiver Kompetenzen mittels Leistungstests, sofern ein standardisiertes Vorgehen bei der Korrektur angewendet wird und gewisse Regeln bei der Korrektur eingehalten werden. Dazu gehört insbesondere eine ausführliche Einarbeitungsphase, in der die korrigierenden Lehrpersonen einen gemeinsamen Beurteilungsmassstab definieren und die Erwartungen absprechen.

Anhang 1


Analyse Schreibanlass Zürich 2010 Originale und Doppelte neWed Feb 10 06:01 2010
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
Terms in the Model (excl Step terms)
+rater
+thema
+item
6
5
4
3
2
1
0
-1
-2
-3
16
17
12
8
6
18
11
9
14
4 5 13
2
1
3 10
15
7
Each 'X' represents 6.4 cases

Anhang 2

```
================================================================================
Analyse Schreibanlass Zürich 2010 Originale und Doppelte neWed Feb 10 06:01 2010
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
================================================================================
TERM 1: rater
--------------------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT             WEIGHTED FIT
---------------                      ----------------------   ----------------------
    rater           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------------------------
 1   Rater 1         -0.063   0.013    1.01 ( 0.91, 1.09)  0.3   1.06 ( 0.90, 1.10)  1.3
 2   Rater 2          0.063*  0.013    0.98 ( 0.83, 1.17) -0.2   0.98 ( 0.82, 1.18) -0.2
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability Not Applicable
Chi-square test of parameter equality =      23.06,  df = 1
^ Quick standard errors have been used
================================================================================
TERM 2: thema
--------------------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT             WEIGHTED FIT
---------------                      ----------------------   ----------------------
    thema           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------------------------
 1   Fernseher       -0.021   0.013    1.03 ( 0.90, 1.10)  0.6   1.09 ( 0.89, 1.11)  1.5
 2   Auto             0.021*  0.013    0.94 ( 0.83, 1.17) -0.7   0.91 ( 0.81, 1.19) -1.0
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability Not Applicable
Chi-square test of parameter equality =       2.55,  df = 1
^ Quick standard errors have been used
================================================================================
TERM 3: item
--------------------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT             WEIGHTED FIT
---------------                      ----------------------   ----------------------
     item           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------------------------
 1   A1_2            -0.724   0.034    2.05 ( 0.91, 1.09) 18.3   1.64 ( 0.88, 1.12)  9.1
 2   A1_3            -0.469   0.036    1.67 ( 0.91, 1.09) 12.7   1.52 ( 0.91, 1.09)  9.9
 3   A_2_1           -1.026   0.037    0.98 ( 0.91, 1.09) -0.5   1.14 ( 0.89, 1.11)  2.4
 4   A_2_2           -0.322   0.035    1.46 ( 0.91, 1.09)  9.2   1.35 ( 0.89, 1.11)  5.8
 5   A_2_3_1         -0.430   0.035    0.94 ( 0.91, 1.09) -1.4   1.08 ( 0.89, 1.11)  1.4
 6   A_2_3_2          0.775   0.033    1.06 ( 0.91, 1.09)  1.4   1.10 ( 0.92, 1.08)  2.3
 7   L_3_1           -1.435   0.038    0.97 ( 0.91, 1.09) -0.6   1.08 ( 0.90, 1.10)  1.6
 8   L_3_2            1.084   0.042    0.80 ( 0.91, 1.09) -4.9   1.00 ( 0.82, 1.18)  0.0
 9   L_3_3_1         -0.078   0.035    1.33 ( 0.91, 1.09)  6.7   1.35 ( 0.91, 1.09)  6.8
 10  L_3_3_2         -1.056   0.037    1.16 ( 0.91, 1.09)  3.5   1.16 ( 0.88, 1.12)  2.4
 11  L_3_3_3          0.152   0.035    1.19 ( 0.91, 1.09)  4.2   1.24 ( 0.91, 1.09)  5.0
 12  L_3_4            1.239   0.034    1.01 ( 0.91, 1.09)  0.1   1.05 ( 0.92, 1.08)  1.3
 13  L_3_5           -0.384   0.038    0.75 ( 0.91, 1.09) -6.1   0.79 ( 0.91, 1.09) -5.1
 14  L_3_6           -0.186   0.036    0.68 ( 0.91, 1.09) -8.1   0.71 ( 0.92, 1.08) -7.6
 15  L_3_7           -1.085   0.034    0.98 ( 0.91, 1.09) -0.3   1.20 ( 0.90, 1.10)  3.6
 16  G_4_1            1.893   0.034    0.70 ( 0.91, 1.09) -7.6   0.71 ( 0.91, 1.09) -7.4
 17  G_4_2            1.410   0.034    0.97 ( 0.91, 1.09) -0.6   1.02 ( 0.91, 1.09)  0.4
 18  G_4_3            0.640*  0.148    0.80 ( 0.91, 1.09) -5.0   0.80 ( 0.91, 1.09) -4.8
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.999
Chi-square test of parameter equality =   12295.59,  df = 17,  Sig Level = 0.000
^ Quick standard errors have been used
```