Universität Zürich
Institut für Bildungsevaluation

Institut für Bildungsevaluation
Assoziiertes Institut
der Universität Zürich

Entwicklung und Durchführung der Schlussprüfungen an der Weiterbildungsschule Basel-Stadt

Überprüfung des Korrekturverfahrens
im Lehrplanbereich «Texte schreiben»

Florian Keller
Zürich, Oktober 2010

Institut für Bildungsevaluation
Assoziiertes Institut der Universität Zürich
Wilfriedstrasse 15
8032 Zürich

Tel. 043 268 39 60
Fax 043 268 39 67

E-Mail: florian.keller@ibe.uzh.ch

Inhalt

1 Ausgangslage

Am Ende der obligatorischen Schulzeit werden an der Weiterbildungsschule Basel-Stadt (WBS) Schlussprüfungen durchgeführt. Mit den Schlussprüfungen werden die Fähigkeiten der Schülerinnen und Schüler in den Fächern Deutsch und Mathematik anhand eines Leistungstests geprüft und benotet.

Der Mathematiktest prüft die vier Lehrplanbereiche «Zahlen und Zahlenoperationen», «Funktionen und Gleichungen», «Geometrie» sowie «Stochastik», der Deutschtest die Lehrplanbereiche «Texte verstehen» und «Sprachbetrachtung» sowie den Bereich «Texte schreiben».

Der Lehrplanbereich «Texte schreiben» wird mit einem Schreibanlass geprüft. Dazu werden den Schülerinnen und Schülern drei Themen vorgelegt. Die Schülerinnen und Schüler wählen ein Thema aus und schreiben dazu in 40 Minuten einen Text, der in der Regel ein bis drei Seiten umfasst. Die Texte der Schülerinnen und Schüler werden anschliessend am Institut für Bildungsevaluation (IBE) von drei eigens dafür geschulten Personen – Lehrpersonen und Germanistinnen – anhand eines standardisierten Korrekturrasters bewertet.

Die Bewertung fliesst als Teil des Prüfungsergebnisses in die Deutschnote ein. Ebenso wird der Bereich «Texte schreiben» in den Klassenrückmeldungen als eigenständiger Teilbereich aufgeführt. In den Schlussbericht hingegen, in dem die wichtigsten Ergebnisse der Schlussprüfungen zusammengefasst und die Leistungsentwicklung an der WBS dargestellt werden, findet der Bereich «Texte schreiben» keinen Eingang. Die im Schlussbericht rapportierte durchschnittliche Deutschleistung an der WBS wird nur aus den Teilbereichen «Texte verstehen» und «Sprachbetrachtung» berechnet.

Da die Aufsatzthemen jedes Jahr neu gestellt wurden, war es bislang nicht möglich, die Leistungen auf einer einheitlichen Skala zu normieren und so vergleichbar zu machen. Selbstverständlich wäre es möglich, die Kompetenzen der Schülerinnen und Schüler im Bereich «Texte schreiben» über die Link-Items der anderen Bereiche zu schätzen. Da mit dem Schreiben eines Textes – im Gegensatz zu den Bereichen «Texte verstehen» und «Sprachbetrachtung» – nicht reproduktive Kompetenzen, sondern produktive Kompetenzen getestet werden, ist eine Integration der Schreibergebnisse über andere Lehrplanbereiche allerdings nicht sinnvoll gerechtfertigt.

Im Deutschtest 2010 standen für einmal die gleichen drei Aufsatzthemen zur Auswahl wie im Jahr 2009: «Mein schönstes Schulerlebnis», «Was ich von einer guten Freundin/einem guten Freund erwarte» und «Das Auto – Fluch oder Segen unserer Zeit?». Für die Bewertung der Texte wurde 2009 und 2010 der gleiche Kriterienraster eingesetzt. Somit ist es möglich, die Leistungsdaten nach der probabilistischen Testtheorie zu skalieren und (1) die Qualität des Korrekturrasters zu überprüfen, (2) die Validität der Korrekturen zu kontrollieren sowie erstmals auch (3) die Leistungen im Bereich «Texte schreiben» zwischen den beiden Prüfungsjahren zu vergleichen.

2 Vorgehen

Die Qualität des Korrekturrasters wird mit einer Analyse der Beurteilungskriterien (Itemanalyse) geprüft. Dazu werden in einem ersten Schritt die Ergebnisse im Bereich «Texte schreiben» aller Schülerinnen und Schüler der Jahre 2009 und 2010 in einem Datensatz zusammengefügt. Danach werden die Daten gemeinsam skaliert und eine Itemanalyse wird durchgeführt. Mit der Itemanalyse werden Schwierigkeit und Trennschärfe sowie die Eindimensionalität des Korrekturrasters überprüft.

In einem zweiten Schritt werden die Daten getrennt für die Prüfungsjahre 2009 und 2010 skaliert und die Schwierigkeitsparameter der Items in jedem Jahr getrennt berechnet. Der Vergleich der beiden Schwierigkeitsparameter pro Item erlaubt die Prüfung der Stichprobenunabhängigkeit. Items, die die empirischen Vorgaben wie Eindimensionalität oder Stichprobenunabhängigkeit nicht erfüllen, werden aus den weiteren Analysen ausgeschlossen. Mit einem Multi-Facetten-Modell werden anschliessend die Strenge der Rater sowie die Schwierigkeit der einzelnen Aufsatzthemen berechnet.

Um die Leistungsentwicklung zwischen den Jahren 2009 und 2010 zu beschreiben, werden die Leistungsdaten aller Schülerinnen und Schüler gemeinsam skaliert und die individuellen Fähigkeitsparameter berechnet. Anhand der Fähigkeitsparameter können die Mittelwerte und die Leistungsverteilungen der beiden Prüfungsjahrgänge verglichen werden.

Die Analysen werden mit der Software ACER Conquest durchgeführt.

3 Stichprobe

Tabelle 1 zeigt eine Übersicht über die Anzahl Schülerinnen und Schüler, die an den Schlussprüfungen 2009 beziehungsweise 2010 teilgenommen und einen beurteilbaren Text abgegeben haben. Insgesamt stehen Daten von 1656 Schülerinnen und Schülern für die Analysen zur Verfügung.

Tabelle 1: Anzahl Schülerinnen und Schüler nach Leistungszug und Testjahr

	2009	2010
Anzahl Schülerinnen und Schüler in Regelklassen des **A-Zugs**	356	332
Anzahl Schülerinnen und Schüler in Regelklassen des **E-Zugs**	408	409
Schülerinnen und Schüler in anderen Klassentypen	70	81
Total	834	822

4 Itemanalysen

4.1 Anzahl erreichter Punkte pro Item

Die Texte der Schülerinnen und Schüler wurden anhand eines standardisierten Kriterienrasters beurteilt. Der Kriterienraster bestand aus zehn inhaltlichen und aus zehn formalen Kriterien (Items), die sich teilweise je nach gewähltem Aufsatzthema unterscheiden. Insgesamt kamen 43 verschiedene Kriterien zum Einsatz. Jedes Kriterium konnte mit 0 Punkten («nicht erfüllt»), mit 1 Punkt («teilweise erfüllt») oder mit 2 Punkten («vollständig erfüllt») bewertet werden[1].

Tabelle 2 zeigt für jedes *inhaltliche* Kriterium den Anteil an Schülerinnen und Schülern, die mindestens 1 Punkt («p corr 1») bzw. 2 Punkte («p corr 2») erreichten. Zudem ist ersichtlich, wie viele Texte («N») und welche Themen («Thema 1», «Thema 2» oder «Thema 3») mit dem jeweiligen Kriterium beurteilt wurden.

Wie Tabelle 2 zeigt, ereichen bei nahezu allen Kriterien rund 90 Prozent oder mehr Schülerinnen und Schüler mindestens einen Punkt. Die im Kriterienraster vorgesehene Differenzierung zwischen «teilweise erfüllt» (1 Punkt) und «vollständig erfüllt» (2 Punkte) kann scheinbar an den Texten nicht nachvollzogen werden und wird von den Ratern kaum genutzt. Um die Beurteilungen der Texte durch die Rater einheitlicher und wohl auch effizienter zu gestalten, wären deshalb durchaus auch dichotome Einschätzungen («nicht oder nur teilweise erfüllt» und «ganz erfüllt») möglich.

Vergleichsweise schwierig mindestens einen Punkt zu erreichen ist es bei den Kriterien T2.5, T2.6 und T2.7 sowie T3.7, mit denen inhaltliche Textbausteine beurteilt werden. Bei diesen Kriterien erreichen weniger als 75 Prozent der Schülerinnen und Schüler mindestens einen Punkt. Daneben ist es auch bei den Kriterien T1.8 «Inneres Erleben wird intensiv miteinbezogen» und T1.9 «Die Geschichte weist Spannungsmomente und einen Höhepunkt auf», die auf den ersten Blick zwar schwierig zu quantifizieren, dafür aber wohl einfacher mehrstufig zu beurteilen sind, schwierig, mindestens einen Punkt zu erreichen.

Am schwierigsten zwei Punkte zu erreichen ist es bei Kriterium T3.7 «Gegenargument wird entkräftet». Nur 20 Prozent der Schülerinnen und Schüler erreichen das Maximum von 2 Punkten. Der grösste Anteil von Schülerinnen und Schülern mit 2 Punkten findet sich bei Kriterium T2.1 «Ausführungen passen zum gestellten Thema». 95 Prozent der Schülerinnen und Schüler erreichten bei diesem Kriterium 2 Punkte.

[1] Der Kriterienraster zur Beurteilung der Texte befindet sich im Anhang.

Tabelle 2: Anteil Schülerinnen und Schüler, die mindestens 1 Punkt bzw. 2 Punkte erreichten pro Kriterium (Inhaltliche Kriterien)

Item	Inhalt	Thema 1	Thema 2	Thema 3	N	pcorr 1	pcorr 2
T1.1	Geschichte passt zum gestellten Thema				853	96%	82%
T1.2	Titel weckt Interesse				853	91%	41%
T1.3	Text fokussiert auf ein einziges Ereignis				853	98%	80%
T1.4	wer, wann, wo?				853	99%	64%
T1.5	Ereignis im Hauptteil vollständig				853	98%	79%
T1.6	Schluss rundet Geschichte ab				853	92%	58%
T1.7	Handlungsschritte sind verbunden				853	98%	80%
T1.8	Inneres Erleben				853	79%	29%
T1.9	Spannungsmomente				853	83%	33%
T1.10	Originalität ist vorhanden				853	90%	33%
T2.1	Ausführung passen zum Thema				803	99%	95%
T2.2	Aufbau enthält die verlangten Inhalte				410	96%	57%
T2.3	Erster Teil enthält ganze Aussagen				410	94%	67%
T2.4	Zweiter Teil enthält ganze Aussagen				410	93%	67%
T2.5	Dritter Teil enthält Situationsbeschreibung				410	72%	51%
T2.6	Vierter Teil enthält Erklärung				410	72%	56%
T2.7	Schlussfolgerung				410	66%	32%
T2.8	Gedanken sinnvoll verbunden				410	97%	59%
T2.9	Sachliche Aussagen sind richtig				803	100%	88%
T2.10	Originalität ist erkennbar				803	89%	34%
T3.2	Einleitung führt zum Thema hin				393	86%	49%
T3.3	Mindestens zwei Argumente und mindestens ein Gegenargument				393	99%	88%
T3.4	Erstes Argument: Aussage und Beleg				393	99%	72%
T3.5	Zweites Argument: Aussage und Beleg				393	92%	59%
T3.6	Gegenargument: Aussage und Beleg				393	96%	69%
T3.7	Gegenargument wird entkräftet				393	50%	20%
T3.8	Schlussfolgerung mit Stellungnahme				393	90%	59%

Tabelle 3 zeigt für jedes *formale* Kriterium, den Anteil an Schülerinnen und Schülern, die mindestens 1 Punkt («p corr 1») bzw. 2 Punkte («p corr 2») erreichten. Zudem ist ersichtlich, wie viele Texte («N») und welche Themen («Thema 1», «Thema 2» bzw. «Thema 3») mit dem jeweiligen Kriterium beurteilt wurden.

Tabelle 3: Anteil Schülerinnen und Schüler, die mindestens 1 Punkt bzw. 2 Punkte erreichten pro Kriterium (Formale Kriterien)

Item	Inhalt	Thema 1	Thema 2	Thema 3	N	pcorr 1	pcorr 2
T1.11	Satzbau vollständig und enthält Nebensätze				853	98%	58%
T1.12	Text enthält Metapher usw.				853	22%	6%
T1.13	Wortschatz				1656	98%	56%
T1.14	keine Leerformeln				853	97%	52%
T1.15	Verbformen: Präteritum				853	99%	72%
T1.16	Vorzeitigkeit = Plusquamperfekt				853	26%	8%
T1.17	Korrektes Geschlecht und richtige Fallformen				1656	94%	71%
T1.18	Satzschlusszeichen, Kommas und Redezeichen				853	92%	41%
T1.19	Rechtschreibung im Verhältnis zum Wortschatz				1656	95%	60%
T1.20	Grossschreibung beherrscht				1656	96%	71%
T2.11	Satzbau vollständig, korrekt, richtig abgegrenzt				803	97%	68%
T2.12	Text enthält Nebensätze				803	90%	60%
T2.13	Sätze untereinander verbunden				803	98%	69%
T2.15	Korrekte Verbformen				803	99%	87%
T2.17	Satzschlusszeichen, Kommas richtig				803	94%	44%
T2.20	Bildet Abschnitte und achtet auf Darstellung				803	74%	56%

Nahezu alle formalen Kriterien werden von über 90 Prozent der Schülerinnen und Schüler zumindest teilweise erfüllt (1 Punkt). Am wenigsten Punkte werden bei jenen Kriterien erreicht, mit denen ganz präzis einzelne formale oder grammatikalische Aspekte beurteilt werden wie T1.12 «Text enthält Vergleiche, Metapher oder andere Stilmittel» oder T1.16 «Vorzeitigkeit im Plusquamperfekt». Letztlich wären aber auch bei den formalen Kriterien dichotome Einschätzungen («nicht oder nur teilweise erfüllt» und «ganz erfüllt») denkbar bzw. vorzuziehen.

4.2 Mittlere Trennschärfe der Items

Die Trennschärfe zeigt, 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. Items mit einem Trennschärfekoeffizienten von grösser als $r_{it} = 0.30$ werden als hinreichend trennscharf beurteilt.

Tabelle 4 zeigt die mittleren Trennschärfen für alle inhaltlichen Kriterien. Alle Items mit Ausnahme der Kriterien T1.1 und T1.4 haben eine Trennschärfe von mehr als 0.30. Am besten zwischen guten und schwachen Schülerinnen und Schülern trennt Kriterium T2.8 «Gedanken sind sinnvoll verbunden» mit einer sehr hohen Trennschärfe von $r_{it} = 0.74$.

Tabelle 4: Mittlere Trennschärfe pro Kriterium (Inhaltliche Kriterien)

Item	Inhalt	Trennschärfe
T1.1	Geschichte passt zum gestellten Thema	0.23
T1.2	Titel weckt Interesse	0.44
T1.3	Text fokussiert auf ein einziges Ereignis	0.39
T1.4	wer, wann, wo?	0.28
T1.5	Ereignis im Hauptteil vollständig	0.48
T1.6	Schluss rundet Geschichte ab	0.52
T1.7	Handlungsschritte sind verbunden	0.50
T1.8	Inneres Erleben	0.52
T1.9	Spannungsmomente	0.62
T1.10	Originalität ist vorhanden	0.66
T2.1	Ausführung passen zum Thema	0.37
T2.2	Aufbau enthält die verlangten Inhalte	0.65
T2.3	Erster Teil enthält ganze Aussagen	0.47
T2.4	Zweiter Teil enthält ganze Aussagen	0.51
T2.5	Dritter Teil enthält Situationsbeschreibung	0.59
T2.6	Vierter Teil enthält Erklärung	0.54
T2.7	Schlussfolgerung	0.52
T2.8	Gedanken sinnvoll verbunden	0.74
T2.9	Sachliche Aussagen sind richtig	0.38
T2.10	Originalität ist erkennbar	0.64
T3.2	Einleitung führt zum Thema hin	0.48
T3.3	Mindestens zwei Argumente und mindestens ein Gegenargument	0.46
T3.4	Erstes Argument: Aussage und Beleg	0.50
T3.5	Zweites Argument: Aussage und Beleg	0.53
T3.6	Gegenargument: Aussage und Beleg	0.55
T3.7	Gegenargument wird entkräftet	0.41
T3.8	Schlussfolgerung mit Stellungnahme	0.52

Tabelle 5 zeigt die mittleren Trennschärfen für alle formalen Kriterien. Die formalen Kriterien trennen noch stärker als die inhaltlichen Kriterien zwischen den Schülerinnen und Schülern. Alle Items haben eine Trennschärfe von mehr als 0.30. Die geringste Trennschärfe hat Kriterium T1.16 «Vorzeitigkeit im Plusquamperfekt» (r_{it} = 0.32), die höchste Trennschärfe Kriterium T2.11 «Satzbau vollständig, korrekt, richtig abgegrenzt» (r_{it} = 0.66).

Tabelle 5: Mittlere Trennschärfe pro Kriterium (Formale Kriterien)

Item	Inhalt	Trennschärfe
T1.11	Satzbau vollständig und enthält Nebensätze	0.57
T1.12	Text enthält Metapher usw.	0.45
T1.13	Wortschatz	0.62
T1.14	keine Leerformeln	0.56
T1.15	Verbformen: Präteritum	0.40
T1.16	Vorzeitigkeit im Plusquamperfekt	0.32
T1.17	Korrektes Geschlecht und Fallformen	0.52
T1.18	Satzschlusszeichen, Kommas und Redezeichen	0.38
T1.19	Rechtschreibung im Verhältnis zum Wortschatz	0.45
T1.20	Grossschreibung beherrscht	0.46
T2.11	Satzbau vollständig, korrekt, richtig abgegrenzt	0.66
T2.12	Text enthält Nebensätze	0.58
T2.13	Sätze untereinander verbunden	0.63
T2.15	Korrekte Verbformen	0.51
T2.17	Satzschlusszeichen, Kommas richtig	0.47
T2.20	Bildet Abschnitte	0.41

4.3 Eindimensionalität des Tests

Um sicher zu gehen, dass mit den inhaltlichen und formalen Kriterien die gleichen Fähigkeiten, nämlich die Fähigkeit «Texte schreiben» beurteilt werden, muss die Eindimensionalität des Kriterienrasters überprüft werden. Diese Eindimensionalität ist eine Voraussetzung, um die Daten überhaupt nach der probabilistischen Testtheorie skalieren zu können.

Dazu wurden die Daten zweidimensional skaliert. Eine Dimension umfasste alle Items, mit denen inhaltliche Kriterien beurteilt wurden, eine zweite Dimension umfasste alle Items, mit denen formale Kriterien beurteilt wurden. Die Korrelation zwischen der Dimension «Inhalt» und der Dimension «Form» beträgt r = 0.72. Das heisst, die beiden Dimensionen hängen so stark zusammen, dass von einer eindimensionalen Beurteilung ausgegangen werden kann. Alle zur Beurteilung der Texte eingesetzten Kriterien messen hinreichend die gleiche Fähigkeit.

4.4 Modellkonformität der Items

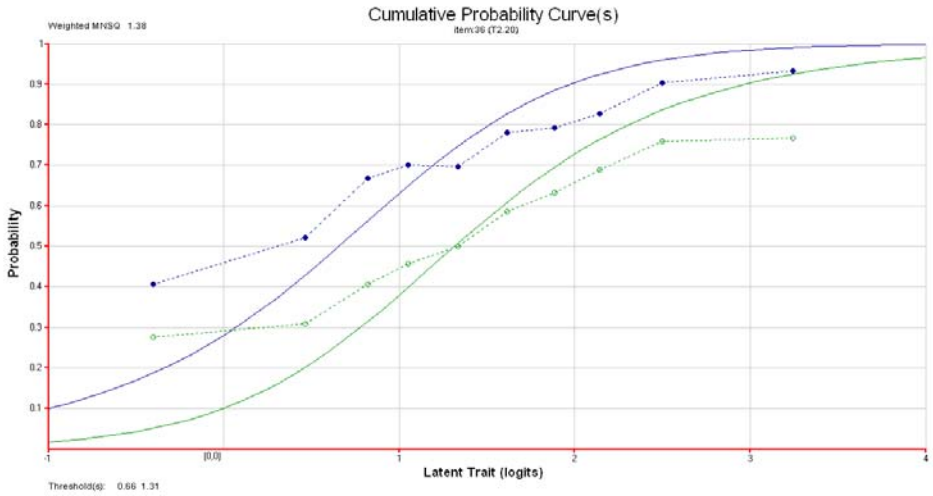
Die Modellkonformität der Items kann anhand der Weighted Infits (MNSQ) überprüft werden. Als Kriterien für eine Verletzung der Modellkonformität der Items gelten MSNQ > 1.20 oder MNSQ < 0.80 sowie auffällig grosse positive T-Werte. Zudem werden die Item Characteristic Curves (ICC) aller Items überprüft. Mit den ICC wird die Beziehung zwischen der Schwierigkeit eines Kriteriums, der Fähigkeit der Schülerinnen und Schülern und der Lösungswahrscheinlichkeit grafisch dargestellt. Für jedes Kriterium und für jeden Bewertungpunkt können zwei Kurven gezeichnet werden. Eine Kurve beschreibt die vom Modell vorhergesagten Werte (durchgezogene Linie), die zweite Kurve beschreibt die empirisch beobachteten Werte (gestrichelte Linie). Weichen die beiden Kurven stark voneinander ab, so ist das betreffende Kriterium nicht modellkonform.

Tabelle 6: Liste der nicht modellkonformen Items

Item	**Inhalt**	N	Delta	**MNSQ**	**T-Value**
T2.20	Bildet Abschnitte	803	0.985	**1.38**	**7.4**

Wie Tabelle 6 zeigt, ist nur ein Item (Item T2.20) nicht modellkonform und kann nicht raschskaliert werden. Dies illustriert auch die ICC dieses Kriteriums (Abbildung 1). Für die weiteren Analysen wird Item T2.20 deshalb ausgeschlossen.

Abbildung 1: IC-Kurve des Kriteriums T2.10 «Bildet Abschnitte und achtet auf die Darstellung»

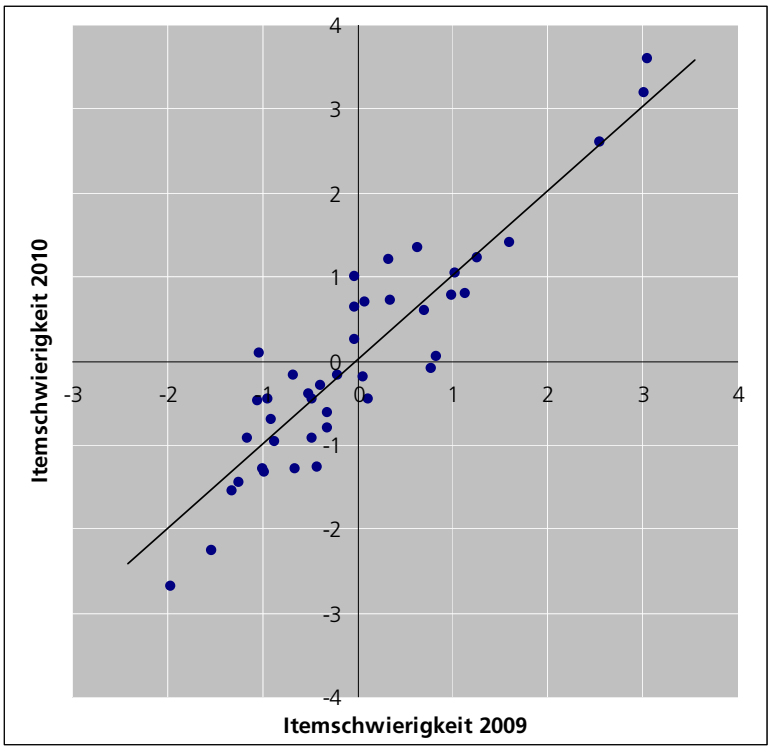


4.5 Stichprobenunabhängigkeit

Um die Stichprobenunabhängigkeit der Beurteilungskriterien zu prüfen, wurden die Daten der Schlussprüfung 2009 und die Daten der Schlussprüfung 2010 noch einmal getrennt skaliert. Die Gegenüberstellung der berechneten Itemschwierigkeiten der beiden Prüfungsjahre in einem zweidimensionalen Koordinatensystem ermöglicht einen grafischen Test: Items, die den Annahmen des Rasch-Modells entsprechen, liegen in dieser Darstellung auf der 45-Grad-Geraden des Koordinatensystems. Items mit grossen Abweichungen zwischen den beiden Schwierigkeitsparametern können beispielsweise auf unterschiedliches Korrekturverhalten bei den einzelnen Items hindeuten. Sie können nicht raschskaliert werden.

Abbildung 2 zeigt das Ergebnis dieses Vergleichs. Jeder schwarze Punkt steht für eine Testaufgabe. Die Position der Punkte ist durch die Itemschwierigkeiten der Prüfungsergebnisse 2009 und durch die Itemschwierigkeiten der Prüfungsergebnisse 2010 definiert. Aufgaben, die weit von der Winkelhalbierenden durch den Nullpunkt entfernt liegen, zeigen zwischen den Testjahren grosse Veränderungen in der relativen Schwierigkeit.

Abbildung 2: Vergleich der Aufgabenschwierigkeiten der Items in der Schlussprüfung 2009 und in der Schlussprüfung 2010



Wie Abbildung 2 zeigt, unterscheiden sich die Schwierigkeiten einiger Items relativ stark zwischen den Prüfungsjahren. Dies war zu erwarten angesichts der verschiedenen Rater,

die die Texte 2009 und 2010 korrigierten. Die durchschnittliche Abweichung der Itemschwierigkeiten zwischen den beiden Prüfungen beträgt 0.403 Logits.

Tabelle 7 zeigt eine Liste der Items, deren Schwierigkeitsparameter zwischen den Schlussprüfungen 2009 und 2010 um mehr als 0.6 Logits abweichen. In der vierten Spalte sind die Schwierigkeitsparameter der Items aus dem Jahr 2009 («Delta 09»), in der fünften Spalte die Schwierigkeitsparameter der Items aus dem Jahr 2010 («Delta 10»). In der sechsten Spalte ist die Differenz zwischen den beiden Schwierigkeitsparametern aufgeführt.

Tabelle 7: Items mit Abweichungen von mehr als 0.6 Logits zwischen den Schwierigkeitsparametern 2009 und den Schwierigkeitsparametern 2010

Item	**Inhalt**	N	Delta 09	Delta 10	**Diff**
T2.4	Zweiter Teil enthält ganze Aussagen	410	-1.029	0.083	-1.112
T1.10	Originalität ist vorhanden	853	-0.030	0.995	-1.025
T2.10	Originalität ist erkennbar	803	0.329	1.206	-0.877
T2.12	Text enthält Nebensätze	803	0.771	-0.092	0.863
T3.4	Erstes Argument: Aussage und Beleg	393	-0.426	-1.270	0.844
T1.2	Titel weckt Interesse	853	0.829	0.059	0.770
T2.1	Ausführung passen zum Thema	803	-1.528	-2.259	0.731
T2.9	Sachlichen Aussagen sind richtig	803	-1.960	-2.681	0.721
T1.9	Spannungsmomente	853	0.627	1.340	-0.713
T2.17	Satzschlusszeichen, Kommas richtig	803	-0.023	0.634	-0.657
T1.3	Text fokussiert auf ein einziges Ereignis	853	-0.652	-1.284	0.632
T1.18	Satzschlusszeichen, Kommas und Redezeichen	853	0.085	0.697	-0.612

Es gibt sowohl unter den inhaltlichen wie auch unter den formalen Kriterien Items mit stark unterschiedlicher Schwierigkeit zwischen den Schlussprüfungen 2009 und 2010. Besonders grosse Differenzen gibt es bei eher subjektiven Einschätzungen wie der «Originalität», aber auch bei tendenziell objektivierbaren Kriterien wie T2.12 «Text enthält Nebensätze» oder den Kriterien zu den Satzzeichen.

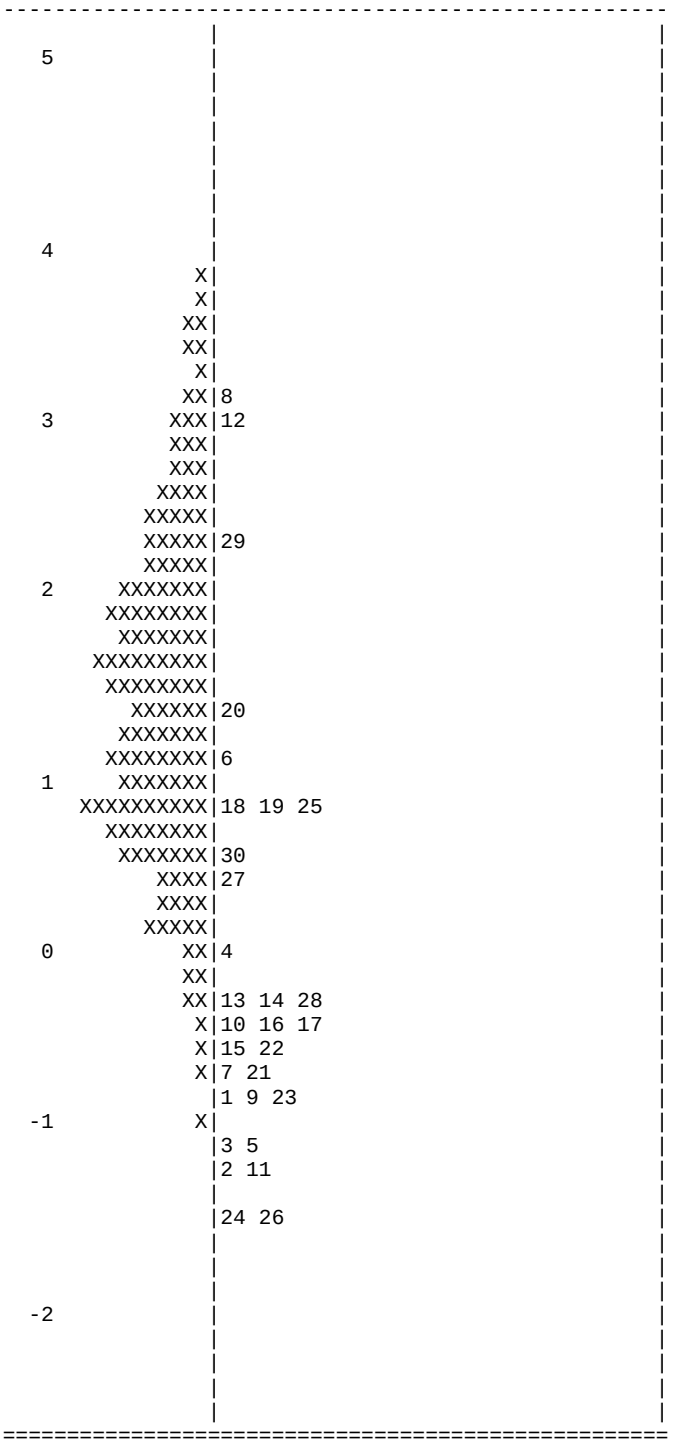
Alle Items mit Differenzen von mehr als 0.6 Logits zwischen den Prüfungsjahren werden aus den Analysen ausgeschlossen. Damit können noch 30 Items skaliert werden. Davon werden 7 Items bei verschiedenen Themen identisch eingesetzt und können als Link-Items verwendet werden.

4.6 Schwierigkeit der Beurteilungskriterien

Abbildung 3 zeigt die Verteilung der Testaufgaben nach Schwierigkeitsgrad. Das einfachste Kriterium (Nummer 24: «Korrekte Verbformen») hat eine mittlere Schwierigkeit von –1.54, das schwierigste Kriterium (Nummer 8: «Text enthält Metapher») hat eine

mittlere Schwierigkeit von 3.15. Es gibt deutlich weniger Kriterien, die hohe Fähigkeiten prüfen als solche, die einfache Fähigkeiten prüfen.

Abbildung 3: Verteilung der Schwierigkeitsparameter der Items und der Fähigkeitsparameter der Schülerinnen und Schüler



In Abbildung 4 ist für jedes Item die Schwierigkeit dargestellt, um mindestens 1 Punkt (Threshold 1) bzw. 2 Punkte zu erreichen (Threshold 2). Wie Abbildung 4 zeigt, sind insbesondere die ersten Thresholds (zu) einfach. Sie prüfen Fähigkeiten, die an der WBS von nahezu allen Schülerinnen und Schülern erreicht werden.

Abbildung 4: Verteilung der Thresholds der Items und der Fähigkeitsparameter der Schülerinnen und Schüler

```
   5            |
                |
                |
                |
                |
                |
                |
                |
   4            |
              X|
              X|8.2
             XX|
             XX|12.2
              X|
             XX|
   3         XXX|29.2
             XXX|
             XXX|
            XXXX|8.1
           XXXXX|12.1
           XXXXX|6.2
           XXXXX|
   2     XXXXXXX|20.2
        XXXXXXXX|
         XXXXXXX|25.2 29.1
       XXXXXXXXX|
        XXXXXXXX|
          XXXXXX|27.2 30.2
         XXXXXXX|10.2 18.2
        XXXXXXXX|9.2 19.2
   1     XXXXXXX|4.2 7.2 14.2 16.2
    XXXXXXXXXX|21.2 28.2
        XXXXXXXX|2.2 20.1
         XXXXXXX|17.2 22.2 23.2
            XXXX|13.2 15.2 18.1 19.1
            XXXX|11.2
           XXXXX|
   0           XX|6.1 25.1
               XX|3.2 5.2
               XX|1.2 30.1
                X|
                X|26.2 27.1
                X|24.2
                 |
  -1             X|4.1
                  |13.1
                  |28.1
                  |1.1 17.1
                  |14.1
                  |15.1
                  |22.1
                  |5.1 16.1
  -2              |
                  |3.1 10.1 21.1
                  |7.1 23.1
                  |24.1 26.1
                  |2.1 9.1 11.1
------------------------------------------------------------------------------
Each 'X' represents  10.8 cases
The labels for thresholds show the levels of item, and step, respectively
```

5 Validität der Beurteilungen

5.1 Strenge der Rater

Die Texte der Schülerinnen und Schüler wurden jedes Jahr von drei Korrektorinnen und Korrektoren (Rater) bewertet. 2009 wurden die Texte von Rater A, Rater B und Rater C korrigiert, 2010 von Rater D, Rater E und wiederum Rater A. Die Zuordnung der Texte zu den Ratern erfolgte nach dem Zufallsprinzip.

Die Beurteilungen wurden während der Korrekturphase in der Gruppe besprochen und durch Doppelkorrekturen validiert. Ziel war es, eine möglichst hohe Übereinstimmung zwischen den Ratern zu erreichen. Wie Tabelle 7 zeigt, wurde dies ansatzweise erreicht.

Tabelle 7 zeigt für jeden Rater die Anzahl korrigierter Texte, die tiefste sowie die höchste vergebene Bewertung, die durchschnittlich vergebene Punktzahl sowie die Standardabweichung als Mass für die Streuung der abgegebenen Bewertungen.

Tabelle 7: Spannweite, Mittelwert und Varianz der abgegebenen Beurteilungen nach Rater

Jahr	Rater	N	Min	Max	M	SD
2009	Rater A-2009	330	5%	100%	71%	16%
2009	Rater B	250	25%	95%	71%	12%
2009	Rater C	254	38%	100%	78%	14%
2010	Rater A-2010	229	30%	100%	73%	15%
2010	Rater D	287	35%	98%	72%	13%
2010	Rater E	306	0%	100%	76%	17%

Am mildesten beurteilte Rater C die Texte. Sie vergab durchschnittlich 78 Prozent der möglichen Punkte. Am strengsten war Rater B mit durchschnittlich 71 Prozent der möglichen Punkte. Rater A, die sowohl 2009 als auch 2010 dem Korrekturteam angehörte, vergab 2009 71 Prozent und 2010 73 Prozent der möglichen Punkte. Angenommen die Leistungen der Schülerinnen und Schüler haben sich zwischen 2009 und 2010 nicht verändert, so blieb die Strenge ihrer Beurteilungen nahezu konstant.

Die gesamte Bandbreite der möglichen Bewertungen von 0 bis 100 Prozent der Punkte nutzt nur Rater E aus. Die meisten anderen Rater bewerten auch die schwächsten Texte kaum tiefer als mit 30 Prozent der möglichen Punkte. Diese eingeschränkte Varianz der Beurteilungen hängt aber auch mit den eingesetzten Kriterien zusammen, bei denen für jeden abgegebenen Text relativ einfach Punkte zu erreichen waren.

Die durchschnittlichen Bewertungen der Rater liegen scheinbar nahe beieinander in einem Bereich zwischen 71 und 78 Prozent der möglichen Punkte. Die Differenz zwischen dem strengsten und dem mildesten Rater von 7 Prozent der Punkte ist bei einer durchschnittlichen Standardabweichung von rund 15% (d = 0.47) jedoch ein nicht zu

unterschätzender Unterschied. Für die Berechnung der Prüfungsnote in Deutsch wird deshalb die Strenge der Rater ausgeglichen. Dazu wird die Punktzahl von Texten, die von strengen Ratern beurteilt wurden, angehoben und die Punktzahl der Texte, die von milden Ratern beurteilt wurden entsprechend reduziert.

Die Varianz in den Urteilen unterscheidet sich zwischen den Ratern nur schwach. Die grösste Streuung in den Bewertungen haben Rater E (17%) und Rater A-2009 (16%). Beurteilungen, die näher beim Mittelwert liegen, vergaben Rater B (12%) und Rater D (13%).

Tabelle 9 zeigt für jeden Rater die Strenge der Korrektur als Logits auf der WBS-Skala sowie die Modellkonformität der Bewertungen (MNSQ und T-Wert).

Tabelle 9: Strenge und Modellkonformität der Korrekturen

	Rater	N	Strenge (Logit)	Schätz-fehler	MNSQ	T-Value
1	Rater A-2009	330	0.105	0.021	1.14	1.7
2	Rater B	250	0.293	0.022	0.97	-0.3
3	Rater C	254	-0.274	0.023	0.98	-0.2
4	Rater A-2010	229	-0.029	0.023	1.09	0.9
5	Rater D	287	0.077	0.022	0.89	-1.3
6	Rater E	306	-0.171	0.050	1.34	3.6

Die Strenge der Rater variiert zwischen Rater B mit einem Logit von 0.293 und Rater C mit einem Logit von -0.274. Die Spannweite der Beurteilungen beträgt somit rund 0.57 Logits. Dies ist angesichts einer Standardabweichung von rund 1 Logit in der Populationsverteilung eine mittelstarke Differenz. Insgesamt ist die Differenz zwischen den Ratern 2009 (0.57 Logits) doppelt so gross wie 2010 (0.25 Logits). Eine Item Map mit der Strenge der Rater befindet sich im Anhang.

Wie Tabelle 9 zeigt, sind die Bewertungen der Rater hinreichend modellkonform. Einzig die Bewertungen von Rater E haben einen hohen MNSQ-Wert, verbunden mit einem vergleichsweise hohen positiven T-Wert. Das heisst, ihre Bewertungen werden auch von Faktoren beeinflusst, die nicht durch Unterschiede in der Qualität der Texte im Bereich «Texte schreiben» erklärt werden können. Die Bewertungen von Rater C, Rater B, aber auch von Rater D zeigen hingegen nahezu keine nicht modellierten Verzerrungen.

Aufgrund dieser Auswertungen sind die Korrekturen von Rater D und Rater A-2010 am besten. Eher ungenügend sind die Korrekturen von Rater B und Rater C, deren Beurteilungen stark vom Gesamtmittelwert abweichen sowie die Korrekturen von Rater E, die eher zu wenig streng und zu wenig einheitlich (zu wenig trennscharf) korrigiert.

Die Strenge der Rater unterscheidet sich jedoch nicht nur zwischen den Ratern, sondern auch je nach Thema des Textes. Die Tabellen 10 bis 12 zeigen für jedes Thema, wie stark die Rater bei der Beurteilung der Texte von ihrer durchschnittlichen Strenge abweichen.

Tabelle 10: Abweichung der Rater von ihrer durchschnittlichen Strenge: Thema 1 «Schulerlebnis»

	Rater	N	Strenge (Logit)	Schätz-fehler	MNSQ	T-Value
1	Rater A-2009	162	0.010	0.025	1.12	1.0
2	Rater B	150	0.101	0.027	0.82	-1.6
3	Rater C	111	0.020	0.027	1.29	2.0
4	Rater A-2010	114	0.056	0.028	1.04	0.3
5	Rater D	169	0.051	0.026	1.02	0.2
6	Rater E	147	-0.239	0.059	0.88	-1.0

Tabelle 11: Abweichung der Rater von ihrer durchschnittlichen Strenge: Thema 2 «Guter Freund»

	Rater	N	Strenge (Logit)	Schätz-fehler	MNSQ	T-Value
1	Rater A-2009	75	0.080	0.030	1.25	1.4
2	Rater B	58	-0.091	0.033	0.69	-1.8
3	Rater C	63	-0.006	0.032	1.14	0.7
4	Rater A-2010	53	-0.051	0.033	1.13	0.7
5	Rater D	64	-0.260	0.033	0.77	-1.2
6	Rater E	97	0.328	0.072	1.28	1.7

Tabelle 12: Abweichung der Rater von ihrer durchschnittlichen Strenge: Thema 3 «Auto»

	Rater	N	Strenge (Logit)	Schätz-fehler	MNSQ	T-Value
1	Rater A-2009	93	-0.090	0.039	1.21	1.3
2	Rater B	42	-0.010	0.042	0.92	-0.3
3	Rater C	80	-0.014	0.042	0.96	-0.2
4	Rater A-2010	62	-0.006	0.043	1.04	0.3
5	Rater D	54	0.209	0.042	0.71	-1.6
6	Rater E	62	-0.089	0.093	1.38	1.8

Wie sich zeigt, sind Rater C und Rater A-2010 bei der Beurteilung der Texte zu den verschiedenen Themen sehr konsistent. Sie weichen bei keinem Thema statistisch signifikant von ihrer durchschnittlichen Strenge ab. Grössere Unterschiede bei der Beurteilung der einzelnen Themen zeigen insbesondere Rater D und Rater E. Rater D beurteilt die Texte zum Thema «Guter Freund» deutlich milder als die übrigen Texte und die Texte zum Thema «Auto» deutlich strenger als die übrigen Texte. Rater E hingegen beurteilt

die Texte zum Thema «Schulerlebnis» milder und die Texte zum Thema «Guter Freund» strenger.

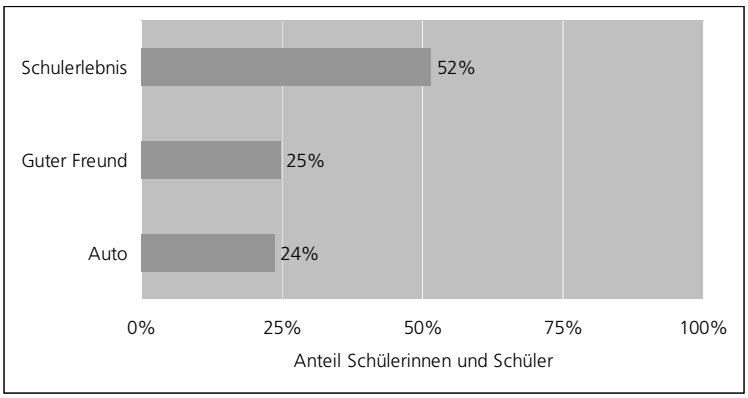
Insgesamt weichen bei der Beurteilung der Texte zum Thema 2 «Guter Freund» vier der sechs Rater statistisch signifikant von ihrer durchschnittlichen Strenge ab. Dies legt die Vermutung nahe, dass entweder die inhaltliche und formale Variation der Texte zum Thema 2 viel breiter und daher eine konsistente Beurteilung schwieriger ist oder dass das Beurteilungsraster für die Beurteilungen der Texte zu diesem Thema nicht adäquat ist.

5.2 Auswahl und Schwierigkeit der Themen

Abbildung 5 zeigt, welches Thema von wie vielen Schülerinnen und Schülern gewählt wurde. Rund die Hälfte der Schülerinnen und Schüler (52%) wählte das Thema «Mein schönstes Schulerlebnis». Je rund ein Viertel der Schülerinnen und Schüler wählten Thema 2 «Was ich von einer guten Freundin/einem guten Freund erwarte» oder Thema 3 «Das Auto – Fluch oder Segen unserer Zeit?».

Dass der grösste Teil der Schülerinnen und Schüler das «Schulerlebnis» als Thema wählte, hängt vermutlich auch mit der Positionierung im Testheft (1. Thema) und der scheinbar freien inhaltlichen Ausrichtung des Themas zusammen. Zudem wurde in der Aufgabenstellung Thema 1 den Schülerinnen und Schülern des A-Zugs empfohlen, während Thema 3 in erster Linie für E-Schülerinnen und -schüler vorgesehen war.

Abbildung 5: Auswahl der Themen



Die unterschiedliche Wahlpräferenz der Themen nach Leistungszug der Schülerinnen und Schüler zeigt Abbildung 6. Das Thema «Schulerlebnis» wird zu 61% von A-Schülerinnen und Schülern und zu 39% von E-Schülerinnen und -schülern gewählt. Thema 2 «Guter Freund» wird zur Hälfte von A- und zur Hälfte von E-Schülerinnen und -schülern gewählt, während Thema 3 «Auto» zu über 90% von E-Schülerinnen und -schülern gewählt wird.

Abbildung 6: Auswahl der Themen nach Leistungszug

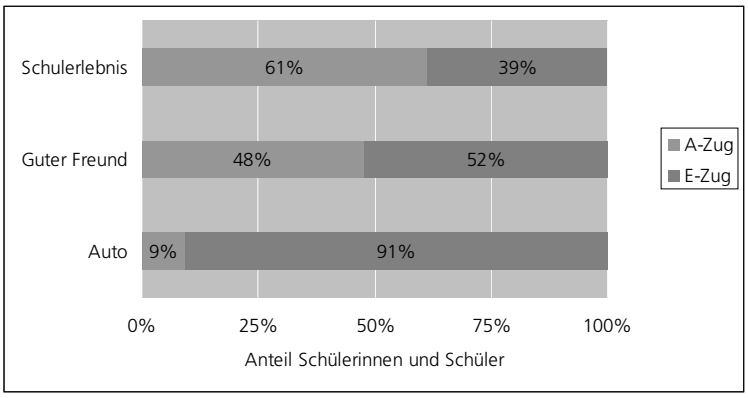
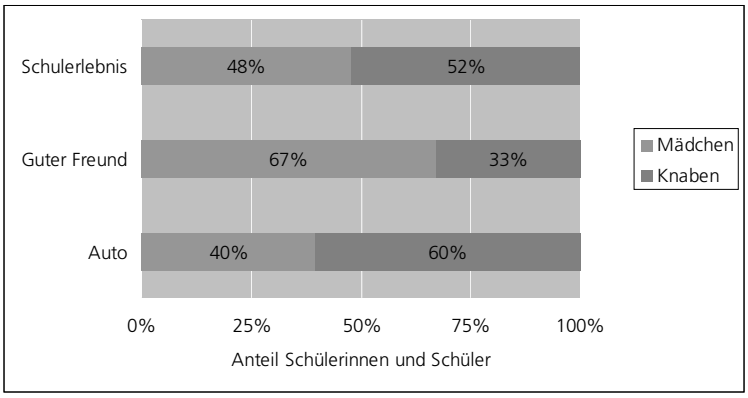


Abbildung 7 zeigt die unterschiedlichen Wahlpräferenzen nach Geschlecht. Thema 1 «Schulerlebnis» wird je zur Hälfte von Mädchen und Knaben gewählt. Thema 2 «Guter Freund» ist eher ein Mädchenthema und Thema 3 «Auto» ist ein Thema, das eher von Knaben gewählt wird.

Abbildung 7: Auswahl der Themen nach Geschlecht



Mit der Raschskalierung ist es nun möglich, die Schwierigkeit der drei Themen unabhängig von der Fähigkeit der Schülerinnen und Schüler und unabhängig von der Strenge der Rater zu schätzen.

Tabelle 13 zeigt das Ergebnis der Skalierung. Das Thema, das am strengsten beurteilt wird, ist Thema 1, das Thema mit der mildesten Beurteilung ist Thema 3. Der Unterschied beträgt 0.718 Logits, was ungefähr 70% einer Standardabweichung in der Population entspricht. Dies ist erstaunlich, ging man bei der Entwicklung der Themen doch davon aus, dass das Thema 1 das einfachste Thema sei. Deshalb wurde Thema 1 auch den Schülerinnen und Schüler des A-Zugs zur Bearbeitung empfohlen.

Tabelle 13: Schwierigkeit und Modellkonformität der Themen

	Thema	N	Schwierigkeit (Logits)	Schätzfehler (Logits)	MNSQ	T-Value
1	Schulerlebnis	853	0.262	0.015	1.02	0.4
2	Guter Freund	410	0.194	0.019	1.01	0.1
3	Auto	393	-0.456	0.024	1.18	2.2

Weshalb Thema 1 am strengsten und Thema 3 am mildesten beurteilt wird, ist nicht einfach zu erklären. Möglicherweise ist das freie Erzählen einer Geschichte sprachlich anspruchsvoller, als angenommen. Vielleicht fliesst beim Thema 1 «Schulerlebnis» auch der subjektiv geprägte Inhalt des Textes stärker in die Beurteilung ein als bei den Themen 2 und 3. Wichtiger ist aber vermutlich, dass die Beurteilungskriterien, die bei allen Texten identisch eingesetzt wurden (Link-Items), je nach Thema unterschiedlich beurteilt werden.

Darauf deuten auch die Resultate der Skalierung hin, wenn zusätzlich noch der Leistungszug der Schülerinnen und Schüler ins Modell aufgenommen wird (Tabelle 14). Thema 1 ist immer noch das schwierigste Thema. Die Unterschiede zwischen den Themen haben sich jedoch stark angenähert[2]. Das lässt vermuten, dass es für Schülerinnen und Schüler, die Thema 3 gewählt haben, schwieriger ist bei den gemeinsamen Kriterien gut beurteilt zu werden als für Schülerinnen und Schüler, die Thema 1 gewählt haben. Ein Vergleich der Schwierigkeit der Themen ist somit streng genommen nicht möglich. Ebenso ist es nicht möglich, die unterschiedlichen Texte empirisch auf eine gemeinsame und einheitlich bewertbare Skala zu bringen.

Tabelle 14: Schwierigkeit und Modellkonformität der Themen nach Kontrolle des Leistungszugs

	Thema	N	Schwierigkeit (Logits)	Schätzfehler (Logits)	MNSQ	T-Value
1	Schulerlebnis	853	0.075	0.016	0.96	-0.8
2	Guter Freund	410	0.013	0.020	1.06	0.8
3	Auto	393	-0.087	0.026	1.25	3.0

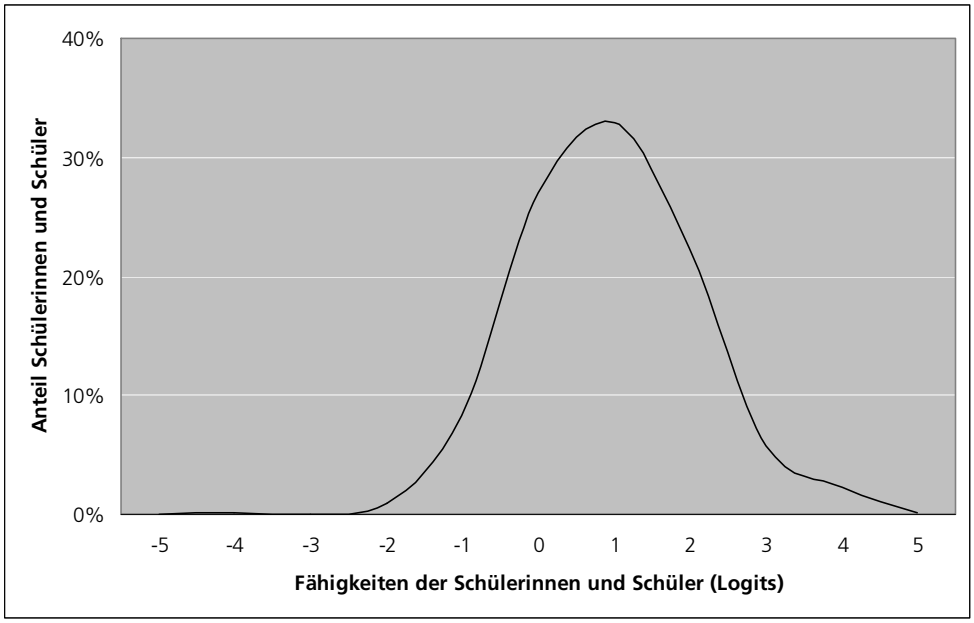
[2] Eine Item-Map mit der Schwierigkeit der einzelnen Themen findet sich im Anhang.

6 Fähigkeiten im Bereich «Texte schreiben»

6.1 Verteilung der Fähigkeiten der Schülerinnen und Schüler

Abbildung 8 zeigt die Verteilung der Fähigkeiten der Schülerinnen und Schüler im Bereich «Texte schreiben» als Logits (Warm Estimates) nach der Korrektur der unterschiedlichen Strenge der Rater.

Abbildung 8: Fähigkeiten der Schülerinnen und Schüler im «Texte schreiben»



Die Fähigkeiten der Schülerinnen und Schüler im Bereich «Texte schreiben» verteilen sich zwischen –5 Logits bis +5 Logits annähernd normal. Die mittlere Fähigkeit liegt bei 1.43 Logits. Die schwächsten 10 Prozent der Schülerinnen und Schüler haben Fähigkeiten von weniger als 0.08 Logits, die besten 10 Prozent haben Fähigkeiten von mehr als 2.81 Logits.

Die durchschnittlichen Fähigkeiten der Schülerinnen und Schüler des A-Zugs liegen bei 0.92 Logits, die durchschnittlichen Fähigkeiten der Schülerinnen und Schüler des E-Zugs bei rund 1.89 Logits.

Mädchen erreichen im Bereich «Texte schreiben» im Durchschnitt Fähigkeiten von 1.55 Logits. Die Fähigkeiten der Knaben liegen 0.25 Logits tiefer bei 1.30 Logits. Dieser Unterschied zwischen Mädchen und Knaben ist statistisch signifikant (Effektstärke d = 0.42).

6.2 Entwicklung der durchschnittlichen Fähigkeiten zwischen 2009 und 2010

Abbildung 9 zeigt die Verteilung der Fähigkeiten der Schülerinnen und Schüler im Bereich «Texte schreiben» getrennt nach den Jahren 2009 und 2010 als Logits. Der kleine Balken in der Mitte gibt an, in welchem Bereich der Mittelwert statistisch gesichert liegt. Die dunkelgrau schattierten Balken links und rechts vom Mittelwert geben den Bereich an, in dem die mittleren 50 Prozent der Fähigkeiten liegen. Zählt man noch den hellgrau schattierten Balken dazu, so erhält man den Bereich, in dem 90 Prozent der Fähigkeiten der Schülerinnen und Schüler liegen.

Abbildung 9: Fähigkeiten der Schülerinnen und Schüler im «Texte schreiben» nach Testjahr

Nachdem die unterschiedliche Strenge der Rater ausgeglichen wurde, liegt der Mittelwert der Fähigkeiten im Jahr 2009 bei 1.44 Logits (SE = 0.04) und im Jahr 2010 bei 1.43 Logits (SE = 0.04). Die durchschnittlichen Fähigkeiten werden damit 2010 um rund 0.01 Logits tiefer eingeschätzt als 2009. Dieser Unterschied ist statistisch jedoch nicht signifikant.

Auch die Verteilung der Fähigkeiten unterscheidet sich zwischen den beiden Testjahren nur schwach. Tendenziell sind die Fähigkeiten im Bereich «Texte schreiben» etwas breiter gestreut als 2009 und der Bereich in dem 90 Prozent der Fähigkeiten liegen, ist 2010 um rund 0.1 Logits tiefer als 2009.

Da die Fähigkeiten in einem Testjahr auch vom Anteil Knaben und vom Anteil E-Schülerinnen und Schüler abhängen, wurden diese Einflussfaktoren mit einer linearen Regression kontrolliert, wobei zugleich die Zugehörigkeit der Schülerinnen und Schüler in die verschiedenen Klassen berücksichtigt wurde (hierarchisch-lineare Regression).

Tabelle 15: Ergebnisse der Regressionsanalyse zur Erklärung der Fähigkeiten im Lehrplanbereich «Texte schreiben»

	Modell 1			Modell 2		
	B	**SE**	**Sig**	**B**	**SE**	**Sig**
Konstante	1.37	0.10	0.000	0.93	0.08	0.000
2010	-0.04	0.14	0.773	0.01	0.09	0.910
Knaben				-0.20	0.05	0.000
E-Schülerinnen und -schüler				1.06	0.09	0.000

Wie Modell 1 in Tabelle 15 zeigt, unterscheiden sich die Mittelwerte im Bereich «Texte schreiben» zwischen den beiden Testjahren nicht statistisch signifikant.

Modell 2 zeigt, dass die Fähigkeiten im Bereich «Texte schreiben» im Jahr 2010 unter Kontrolle des Geschlechts sowie des Leistungszugs um *0.01 Logits* besser waren als im Jahr 2009. Dieser Unterschied ist statistisch aber nicht signifikant. Das heisst, die durchschnittlichen Fähigkeiten im Bereich «Texte schreiben» unterscheiden sich zwischen 2009 und 2010 nur zufällig. Diese Konstanz der Leistungen im «Texte schreiben» entspricht der festgestellten Leistungskontinuität im Fach Deutsch an der WBS insgesamt.

7 Fazit

2009 und 2010 wurden an den Schlussprüfungen der WBS der Bereich «Texte schreiben» genau gleich geprüft. In beiden Jahren wurden den Schülerinnen und Schülern die drei Themen «Mein schönstes Schulerlebnis», «Was ich von einer guten Freundin/einem guten Freund erwarte» und «Das Auto – Fluch oder Segen unserer Zeit?» vorgelegt. Die Texte wurden anhand desselben Kriterienrasters beurteilt. Die insgesamt 1656 Texte wurden in beiden Jahren von je drei Personen (Ratern) korrigiert. Weil der Bereich «Texte schreiben» bei zwei Durchgängen identisch geprüft wurdde, ist es möglich, Kriterien und Korrekturen zu überprüfen und zwischen den Jahre zu vergleichen.

Kriterienraster

Wie die Analysen zeigen, eignet sich das eingesetzte Kriterienraster gut zur Beurteilung der Fähigkeiten im Bereich «Texte schreiben». Die Kriterien sind bis auf wenige Ausnahmen eindimensional und sehr trennscharf. Das heisst, die eingesetzten Kriterien tragen wesentlich dazu bei, die Fähigkeiten der Schülerinnen und Schülern zu bestimmen und zwar so, dass gute Schülerinnen und Schüler eine hohe Punktzahl erreichen und schwache Schülerinnen und Schüler eine tiefe Punktzahl.

Allerdings sind die Kriterien mehrheitlich sehr einfach. Die Fähigkeiten von Schülerinnen und Schüler mit sehr hohen Fähigkeiten im Bereich «Texte schreiben» können nicht mehr genau geschätzt werden und es gibt zu wenige Kriterien, mit denen auch zwischen guten Schülerinnen und Schülern ausreichend differenziert werden kann. Zudem zeigt sich, dass bei fast allen Kriterien nahezu 100 Prozent der Schülerinnen und Schüler mindestens einen Punkt erreichen. Es wäre daher durchaus möglich, die Kriterien mit dichotomer Ausprägung (erreicht/nicht erreicht) ins Kriterienraster aufzunehmen. Eine andere Möglichkeit wäre es, die Kriterien selbst und auch die Abstufungen zwischen den Punkten noch präziser zu definieren. Hier gäbe es allenfalls noch Optimierungsmöglichkeiten für das Kriterienraster.

Rater

Die Korrektur und die Beurteilung der Texte ist insgesamt sehr valide und trennscharf. Sie entspricht den Standards, die an standardisierte Leistungstests gestellt werden. Erwartungsgemäss beurteilen die sechs Rater die Texte unterschiedlich streng und unterschiedlich gut. Aufgrund der Auswertungen sind die Korrekturen von Rater A und Rater D am besten. Mit der Skalierung der Daten ist es jedoch möglich, die Strenge der Rater objektiv zu bestimmen und bei der Notengebung zu berücksichtigen. Dadurch können die Auswirkungen der Rater für die Schülerinnen und Schüler rechnerisch ausgeglichen werden.

Themen

Auch die drei zur Auswahl gestellten Themen unterscheiden sich in ihrer Schwierigkeit. Das Thema «Mein schönstes Schulerlebnis» wird am strengsten, das Thema «Auto – Fluch oder Segen unserer Zeit?» wird am mildesten beurteilt. Dies deutet einerseits darauf hin, dass die formale und inhaltliche Gestaltung von «freien» Themen schwieriger ist als angenommen und dass in diesen Themen normative Aspekte bei der Beurtei-

lung vermutlich stärker ins Gewicht fallen als bei den eher argumentativen Themen. Unter diesen Gesichtspunkten scheint es wenig sinnvoll, den Schülerinnen und Schülern des A-Zugs wie bis anhin das Thema «Mein schönstes Schulerlebnis» zur Bearbeitung zu empfehlen.

Andererseits ist es möglich, dass die Kriterien des Beurteilungsrasters von den Ratern je nach Thema unterschiedlich interpretiert werden. Damit wären die Leistungen der Schülerinnen und Schüler streng genommen nicht mehr vergleichbarFür eine abschliessende Einschätzung des Effekts der Themenwahl auf die Beurteilung der Texte wäre jedoch eine gezielte Untersuchung erforderlich, bei der alle Schülerinnen und Schüler zwei Texte zu unterschiedlichen Themen verfassten, die je von zwei Ratern korrigiert würden.

Leistungsentwickung

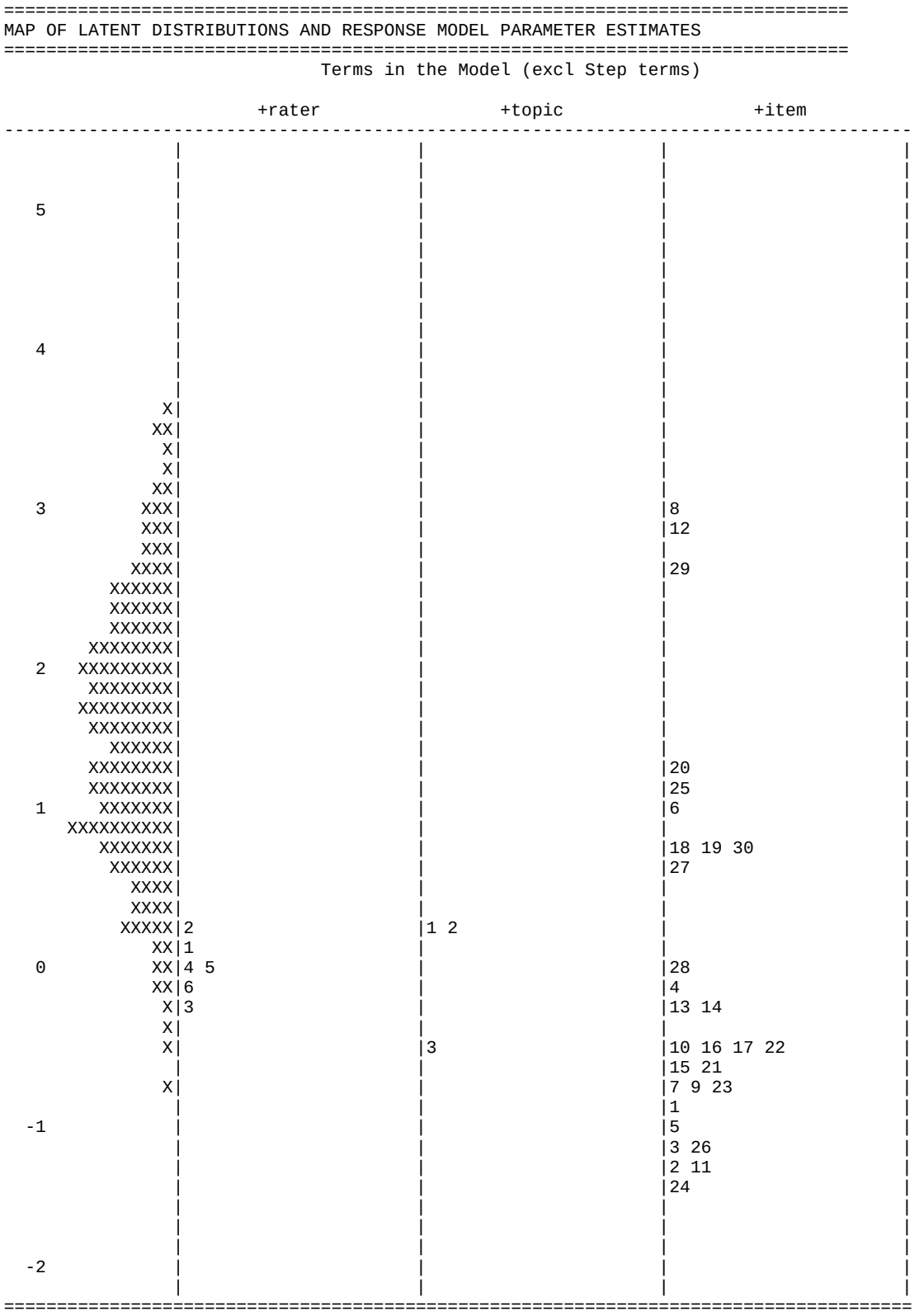
Die durchschnittlichen Fähigkeiten der Schülerinnen und Schüler im Bereich «Texte schreiben» unterscheiden sich nicht zwischen 2009 und 2010. Diese erstaunliche Leistungskonstanz bleibt auch dann bestehen, wenn die unterschiedliche Zusammensetzung der Schülerschaft berücksichtigt wird. Damit wiederspiegelt sich im Teilbereich «Texte schreiben» die allgemeine Leistungskontinuität im Fach Deutsch an der WBS. Dies kann als ein weiteres Zeichen für die Zuverlässigkeit der Korrektur im Bereich «Texte schreiben» interpretiert werden.

Anhang

A) Item-Map der Fähigkeiten der Schülerinnen und Schülern, der Strenge der Rater und der Schwierigkeit der Items (rater+item+item*step)

```
============================================================================
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
============================================================================
                  Terms in the Model (excl Step terms)
                        +rater                  +item
------------------------------------------------------------
                 |                       |                       |
                 |                       |                       |
                 |                       |                       |
   5             |                       |                       |
                 |                       |                       |
                 |                       |                       |
                 |                       |                       |
                 |                       |                       |
                 |                       |                       |
   4             |                       |                       |
                 |                       |                       |
              X  |                       |                       |
              X  |                       |                       |
             XX  |                       |                       |
              X  |                       |                       |
              X  |                       |                       |
             XX  |                       |8                      |
   3        XXX  |                       |12                     |
            XXX  |                       |                       |
            XXX  |                       |                       |
            XXX  |                       |                       |
         XXXXXX  |                       |29                     |
          XXXXX  |                       |                       |
         XXXXXX  |                       |                       |
   2    XXXXXXX  |                       |                       |
       XXXXXXXX  |                       |                       |
       XXXXXXXX  |                       |                       |
     XXXXXXXXXX  |                       |                       |
       XXXXXXXX  |                       |                       |
         XXXXXX  |                       |20                     |
        XXXXXXX  |                       |                       |
       XXXXXXXX  |                       |6                      |
   1   XXXXXXXX  |                       |                       |
    XXXXXXXXXX   |                       |18 19 25               |
        XXXXXXX  |                       |                       |
         XXXXXX  |                       |30                     |
           XXXX  |                       |27                     |
          XXXX   |2                      |                       |
         XXXXX   |1                      |                       |
   0        XX   |4 5                    |4                      |
            XXX  |                       |                       |
             XX  |3 6                    |13 14 28               |
              X  |                       |16 17                  |
              X  |                       |10 22                  |
              X  |                       |7 15 21                |
                 |                       |1 9 23                 |
              X  |                       |                       |
  -1             |                       |5                      |
                 |                       |2 3                    |
                 |                       |11                     |
                 |                       |26                     |
                 |                       |24                     |
                 |                       |                       |
                 |                       |                       |
  -2             |                       |                       |
                 |                       |                       |
============================================================
Each 'X' represents  10.6 cases
============================================================
```

B) Item-Map der Fähigkeiten der Schülerinnen und Schülern, der Strenge der Rater sowie der Schwierigkeit der Themen und der Items (rater+topic+item+item*step)



Korrekturraster

1. MEIN SCHÖNSTES SCHULERLEBNIS

0 Punkt = Merkmal nur in gelegentlichen Ansätzen oder zu weniger als 35 % vorhanden
1 Punkte = Merkmal teilweise oder zu 35%-65% vorhanden
2 Punkte = Merkmal als klare Tendenz oder mehrheitlich zu 65% und mehr vorhanden

Inhaltliche Kriterien

1.1 Die Geschichte passt zum gestellten Thema .....

1.2 Der Titel weckt Interesse, ist gut gewählt .....

1.3 Text fokussiert deutlich auf ein einziges Ereignis .....

1.4 Die Einleitung beantwortet deutlich die Fragen: wer, wo, wann .....

1.5 Das Ereignis ist im Hauptteil vollständig (alle Erzählschritte) nachvollzogen .....

1.6 Der Schluss rundet das Geschehen sinnvoll ab.....

1.7 Die Handlungsschritte sind untereinander verbunden.....

1.8 Inneres Erleben wird intensiv miteinbezogen.....

1.9 Die Geschichte weist Spannungsmomente und einen Höhepunkt auf .....

1.10 Originalität ist vorhanden .....

Formale Kriterien

1.11 Satzbau: vollständig, korrekt, richtig abgegrenzt,
auch Nebensätze enthaltend.....

1.12 Text enthält Vergleiche, Metaphern oder andere Stilmittel.....

1.13 Wortschatz (Verben, nominale adjektivische Bezeichnungen)
ist differenziert, treffend, dem Inhalt angemessen.....

1.14 Vermeidung von Leerformeln: z.B. Wir gingen und
oder Als wir in … angekommen waren, zogen wir uns um etc. .....

1.15 Korrekte Verbformen: Präteritum .....

1.16 Vorzeitigkeit im Plusquamperfekt .....

1.17 Korrektes Geschlecht und richtige Fallformen der nominalen Teile.....

1.18 Satzschlusszeichen, Kommas und Redezeichen richtig.....

1.19 Rechtschreibung im Verhältnis zur Reichhaltigkeit des Wortschatzes.....

1.20 Grossschreibung beherrscht.....

2. WAS ICH VON EINER GUTEN FREUNDIN/EINEM GUTEN FREUND ERWARTE

0 Punkt = Merkmal nur in gelegentlichen Ansätzen oder zu weniger als 35 % vorhanden
1 Punkte = Merkmal teilweise oder zu 35%-65% vorhanden
2 Punkte = Merkmal als klare Tendenz oder mehrheitlich zu 65% und mehr vorhanden

Inhaltliche Kriterien

2.1 Ausführungen passen zum gestellten Thema ..... ☐

2.2 Aufbau enthält die verlangten Inhalte ..... ☐

2.3 Ein erster Teil enthält verständlich ausgeführte Aussagen und Gedanken, nicht nur Aufzählungen und Bruchstücke ..... ☐

2.4 Ein zweiter Teil enthält verständlich ausgeführte Aussagen und Gedanken, nicht nur Aufzählungen und Bruchstücke ..... ☐

2.5 Ein dritter Teil enthält eine Situationsbeschreibung (froh um Freund/in)…. ..... ☐

2.6 Ein vierter Teil enthält eine Erklärung/Begründung (Freundschaft in Brüche) ..... ☐

2.7 Der Text enthält eine überzeugende Schlussfolgerung..... ☐

2.8 Die Gedanken sind sinnvoll verbunden ..... ☐

2.9 Die sachlichen Aussagen sind richtig ..... ☐

2.10 Originalität ist erkennbar, Text vermeidet Allgemeinplätze ..... ☐

Formale Kriterien

2.11 Satzbau, vollständig, korrekt, richtig abgegrenzt..... ☐

2.12 Text enthält auch Nebensätze..... ☐

2.13 Sätze untereinander verbunden, richtige Verweise und Verknüpfungen..... ☐

2.14 Wortschatz (Verben, nominale und adjektivische Bezeichnungen) ist differenziert, treffend, dem Inhalt angemessen, vermeidet Gleichförmigkeit. ..... ☐

2.15 Korrekte Verbformen..... ☐

2.16 Korrektes Geschlecht und richtige Fallformen der nominalen Teile..... ☐

2.17 Satzschlusszeichen, Kommas richtig ..... ☐

2.18 Rechtschreibung im Verhältnis zur Reichhaltigkeit des Wortschatzes..... ☐

2.19 Grossschreibung beherrscht..... ☐

2.20 Bildet Abschnitte und achtet auf Darstellung..... ☐

3. DAS AUTO – FLUCH ODER SEGEN UNSERER ZEIT?

0 Punkt = Merkmal nur in gelegentlichen Ansätzen oder zu weniger als 35 % vorhanden
1 Punkte = Merkmal teilweise oder zu 35%-65% vorhanden
2 Punkte = Merkmal als klare Tendenz oder mehrheitlich zu 65% und mehr vorhanden

Inhaltliche Kriterien

3.1 Ausführungen passen zum gestellten Thema ☐
3.2 Einleitung führt zum Thema hin durch Hinweis auf Aktualität, bezieht noch keine Stellung ☐
3.3 Mindestens zwei Argumente und mindestens ein Gegenargument enthalten ☐
3.4 Erstes Argument besteht aus Aussage und stützendem Beleg ☐
3.5 Zweites Argument ebenfalls ☐
3.6 Gegenargument besteht aus Aussage und stützendem Beleg ☐
3.7 Gegenargument wird entkräftet ☐
3.8 Der Text enthält eine überzeugende Schlussfolgerung mit Stellungnahme ☐
3.9 Die sachlichen Aussagen sind richtig ☐
3.10 Originalität ist erkennbar, Text vermeidet Allgemeinplätze ☐

Formale Kriterien

3.11 Satzbau: vollständig, korrekt, richtig abgegrenzt ☐
3.12 Text enthält auch Nebensätze ☐
3.13 Sätze untereinander verbunden, richtige Verweise und Verknüpfungen ☐
3.14 Wortschatz (Verben, nominale und adjektivische Bezeichnungen) ist differenziert, treffend, dem Inhalt angemessen, vermeidet Gleichförmigkeit. ☐
3.15 Korrekte Verbformen ☐
3.16 Korrektes Geschlecht und richtige Fallformen der nominalen Teile ☐
3.17 Satzschlusszeichen, Kommas richtig ☐
3.18 Rechtschreibung im Verhältnis zur Reichhaltigkeit des Wortschatzes ☐
3.19 Grossschreibung beherrscht ☐
3.20 Bildet Abschnitte und achtet auf Darstellung ☐