Universität Zürich[UZH]

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

# Texte schreiben 2012 – Standardisierte Erfassung von Schreibkompetenzen

Bericht zuhanden der Bildungsdirektion Kanton Zürich

Jeannette Oostlander & Barbara Wespi

Zürich, 12. September 2013

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

Tel: 043 268 39 60
Fax: 043 268 39 67
www.ibe.uzh.ch

Barbara.Wespi@ibe.uzh.ch
Jeannette.Oostlander@ibe.uzh.ch

# Inhalt

# 1 Ausgangslage

Seit dem Schuljahr 2010/11 wird das computergestützte Testsystem „Stellwerk“ im Kanton Zürich in der 8. Klasse flächendeckend eingesetzt. Stellwerk umfasst Tests für die Fachbereiche Deutsch, Englisch, Französisch, Mathematik sowie Natur und Technik. Die Stellwerk-Tests werden ausschliesslich am Computer gelöst, wobei es sich um adaptive Tests handelt, welche sich an die Fähigkeiten der Schülerinnen und Schüler anpassen. Adaptive Tests haben den Vorteil gegenüber dem traditionellen „Papier-Bleistift-Test“, dass die Objektivität der Durchführung gesichert ist und, dass der Computer bei der Korrektur keine Fehler macht. Computergestützte Tests haben allerdings insofern Nachteile, als dass sie nur reproduktive Fähigkeiten testen. Produktive Fähigkeiten wie ausführliche Antworten auf offene Fragen oder ganze Texte können am Computer bis jetzt nicht zufriedenstellend korrigiert und bewertet werden.

Im Dezember 2010 hat der Bildungsrat des Kantons Zürich beschlossen, das Testsystem „Stellwerk“ mit der Erfassung von produktiven Fähigkeiten im Fachbereich Deutsch zu ergänzen. Anhand eines standardisierten Verfahrens soll die Schreibkompetenz der Schülerinnen und Schüler beurteilt werden. Das Institut für Bildungsevaluation hat das Modul „Textproduktion Deutsch“ im Kanton Zürich bereits in den Jahren 2007 bis 2010 als Teil des Pilotprojekts „Neugestaltung 3. Sek“ an insgesamt 4192 Schülerinnen und Schülern erprobt[1]. Im Schuljahr 2012/13 wurde nun der „Stellwerk-Test“ zum ersten Mal mit dem Modul „Texte schreiben“ durchgeführt.

# 2 Durchführung

Die Durchführung des Moduls „Texte schreiben“ fand am 3. und 4. Dezember 2012 statt. Die Teilnahme war für alle Schülerinnen und Schüler der 8. Klasse im Kanton Zürich obligatorisch. Den Schülerinnen und Schülern wurden zwei Themen vorgelegt, wovon sie ein Thema auswählen konnten. Dazu schrieben sie während 60 Minuten einen Text, der in der Regel eine bis drei Seiten umfasste. Insgesamt verfassten 9953 Schülerinnen und Schüler einen Text. Die Texte wurden von den Lehrpersonen kopiert und die Originale dem Institut für Bildungsevaluation zur Beurteilung zugestellt.

Die Texte der Schülerinnen und Schüler wurden anschliessend am Institut für Bildungsevaluation von sieben eigens dafür geschulten Personen – Lehrpersonen, Germanistinnen und Germanisten – anhand eines standardisierten Beurteilungsrasters bewertet (eine detaillierte Darstellung des Beurteilungsrasters befindet sich im Teil 2 dieses Berichtes). Die erreichte Punktzahl im Modul „Texte schreiben“ wurde in die Ergebnisse von „Stellwerk“ integriert, welche von den Lehrpersonen am Computer eingesehen und ausgedruckt werden können. Zugleich wurden die beurteilten Texte an die Lehrpersonen und ihre Schülerinnen und Schüler zurückgesandt. Das ausgefüllte Beurteilungs-

[1] Moser, U. (2007). Standardisierte Erfassung der sprachlichen Kompetenzen im Fachbereich «Texte schreiben». Kurzbericht zuhanden des Pilotprojekts «Neugestaltung des 9. Schuljahrs» und der Projektleitung Sekundarstufe I der Bildungsdirektion des Kantons Zürich.

Moser, U. (2008). Standardisierte Erfassung der sprachlichen Kompetenzen im Fachbereich «Texte schreiben». Kurzbericht zuhanden des Pilotversuchs «Neugestaltung des 9. Schuljahrs» und der Projektleitung Sekundarstufe I der Bildungsdirektion des Kantons Zürich.

Moser, U. & Keller, F. (2009). Standardisierte Erfassung der sprachlichen Kompetenzen im Fachbereich «Texte schreiben». Kurzbericht zuhanden des Pilotprojekts «Neugestaltung 3. Sek» und der Projektleitung der Bildungsdirektion des Kantons Zürich.

Moser, U. (2010). Standardisierte Erfassung der sprachlichen Kompetenzen im Fachbereich «Texte schreiben». Kurzbericht zuhanden des Pilotprojekts «Neugestaltung 3. Sek» und der Projektleitung der Bildungsdirektion des Kantons Zürich.

raster wurde an jeden Aufsatz angeheftet, sodass die Punktevergabe für die Lehrperson und die Schülerinnen und Schüler bei jedem Kriterium einsehbar ist.

## 2.1 Schreibauftrag

Die Ausgangslage für die Entwicklung der Schreibaufträge ist der Lehrplan des Kantons Zürich. Darauf basierend wurde ein Schreibauftrag entwickelt, welcher die Richtziele zum Schreiben möglichst umfänglich abdeckt. Allerdings eignet sich nicht jeder Schreibauftrag gleich gut für eine standardisierte Erfassung der Schreibkompetenzen von Schülerinnen und Schülern. Je offener ein Auftrag formuliert wird, desto schwieriger gestaltet sich die standardisierte Beurteilung. Aus diesem Grund werden Schreibaufträge bevorzugt, welche einen Lebens- beziehungsweise Alltagsbezug haben und die mit klaren Aufgaben verbunden sind.

Mit dem Modul „Texte schreiben" wird angestrebt, die Fähigkeit zu erfassen, Texte verständlich zu formulieren und je nach Zielsetzung adressatengerecht zu schreiben, präzise zu formulieren, überzeugend zu argumentieren oder Sprache ästhetisch ansprechend und kreativ einzusetzen (Harsch, Neumann, Lehmann & Schröder, 2007)[2]

## 2.2 Themen

Im Sinne einer standardisierten schriftlichen Anleitung wurden die Schülerinnen und Schüler aufgefordert, zu einem Sachverhalt Stellung zu nehmen. Die Themen wurden den Schülerinnen und Schülern mittels einer kurzen Einleitung und drei zu beantwortenden Fragen vorgestellt. Die Fragen dienten den Schülerinnen und Schülern als Leitfaden und halfen bei der Strukturierung der Texte. Mit der ersten Frage wurden die Schülerinnen und Schüler dazu aufgefordert, eigene Beobachtungen und Erfahrungen zum Thema zu formulieren und damit die Leserinnen und Leser ins Thema einzuführen. Zwei weitere Fragen zielten auf eine argumentative Stellungnahme zum Thema ab (Darstellung der Vor- und Nachteile). Die Aufgabenstellung entspricht somit einer klassischen Pro-Kontra-Erörterung. Zur Wahl standen zwei Themen:

- Das Leben als Star
- Muss es in der Schule Noten geben?

Die schriftliche Anleitung für beide Themen für den Schreibauftrag ist in den Abbildungen 1 und 2 dargestellt.

[2] Harsch, C., Neumann, A., Lehmann, R. & Schröder, K. (2007). Schreibfähigkeit. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung* (S. 42–62). Weinheim: Belz

Abbildung 1: Schreibauftrag „Das Leben als Star"

**1. Das Leben als Star**

*Stell dir vor, dass du ein Star bist: Ein berühmter Sportler, eine berühmte Sängerin, ein berühmter Musiker, eine berühmte Schauspielerin…*

Verfasse einen Text. Gehe dabei auf folgende Fragen ein:

- Wie sieht dein Leben als Star aus?
- Was sind die Vor- und Nachteile, die ein Leben als Star mit sich bringt?
- Was zählt für dich mehr: die Vorteile oder die Nachteile? Schreibe dazu deine Meinung und begründe sie.

Achte darauf, dass dein Text eine Einleitung, einen Hauptteil und einen Schluss hat.

Abbildung 2: Schreibauftrag „Muss es in der Schule Noten geben?"

**2. Muss es in der Schule Noten geben?**

*In der Schule ist es üblich, die Leistungen mit Noten zu bewerten. Muss das so sein? Findest du das sinnvoll?*

Verfasse einen Text. Gehe dabei auf folgende Fragen ein:

- Welche Erfahrungen machst du mit Noten?
- Was sind die Vor- und Nachteile von Noten in der Schule?
- Was zählt für dich mehr: die Vorteile oder die Nachteile? Schreibe dazu deine Meinung und begründe sie.

Achte darauf, dass dein Text eine Einleitung, einen Hauptteil und einen Schluss hat.

Ebenfalls zum Schreibauftrag gehören Angaben darüber, wie die Texte bewertet werden (siehe Punkt 4 in Abbildung 3). Die ersten beiden inhaltlichen Kriterien beziehen sich auf den expliziten Schreibauftrag. Zum einen muss der kommunikative Auftrag erfüllt werden, zum anderen sollte der Text nach den expliziten Vorgaben strukturiert werden (die Texte müssen eine Einleitung, einen Hauptteil und einen Schluss haben). Darüber hinaus wird den Schülerinnen und Schülern auch mitgeteilt, dass die Kriterien Sprachrichtigkeit („Sind die Sätze korrekt?") und Sprachangemessenheit („Ist die Wortwahl passend?") beurteilt werden.

Abbildung 3: Anleitung zum Schreibauftrag

1. Lies die Informationen zu den zwei Themen genau durch.

2. Wähle eines der zwei Themen aus und verfasse dazu einen Text.

3. Gehe wie folgt vor:

- Schreibe deinen Text auf die ausgeteilten Blätter.
- Schreibe so, dass der Text gut lesbar ist.
- Du darfst Notizpapier zur Vorbereitung gebrauchen.
- Du darfst den Duden oder das Wörterbuch benutzen.

4. Dein Text wird nach folgenden Kriterien bewertet:

- Werden die drei Fragen in deinem Text beantwortet?
- Hat dein Text eine Einleitung, einen Hauptteil und einen Schluss?
- Ist dein Text verständlich?
- Sind die Sätze korrekt?
- Ist die Wortwahl passend?

# 3 Beurteilung der Texte

## 3.1 Beurteilungsraster

Zur Beurteilung der Texte im Rahmen des Moduls „Texte schreiben" wurde ein Beurteilungsraster basierend auf den fünf Basisdimensionen *Inhalt*, *Textaufbau*, *Sprachrichtigkeit*, *Sprachangemessenheit* sowie *Schreibstil und Kreativität* entwickelt. Diese fünf Dimensionen entstanden in Anlehnung an das Zürcher Textanalyseraster von Nussbaumer und Sieber (1994)[3] sowie basierend auf dem Vorgehen von Becker-Mrotzek und Böttcher (2011)[4]. Das „Zürcher Analyseraster" (Nussbaumer & Sieber 1994) erfasst die „sprachsystematische und orthographische Richtigkeit", die „funktionale Angemessenheit", die „ästhetische Angemessenheit" und die „inhaltliche Relevanz". Für die pädagogische Praxis erwies sich das „Zürcher Analyseraster" als zu detailliert und bietet für die inhaltlichen Belange zu wenig Anhaltspunkte für die Beurteilung von Schülertexten (vgl. Neumann, 2007)[5]. Aus diesen Gründen schlagen Becker-Mrotzek und Böttcher (2011) einen Basiskatalog mit zwölf Kriterien vor, verteilt auf die fünf Basisdimensionen „Sprachrichtigkeit", „Sprachangemessenheit", „Inhalt", „Aufbau" und „Schreibprozess". Unter der Basisdimension „Schreibprozess" werden nach Becker-Mrotzek und Böttcher (2011) die folgenden beiden Kriterien zusammengefasst: „Planen/Überarbeiten" (Lässt der Text Planungs- und Überarbeitungsspu-

[3] 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.

[4] Becker-Mrotzek, M. & Böttcher, I. (2011). *Schreibkompetenzen entwickeln und beurteilen.* Berlin: Cornelsen

[5] Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11.* Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen. Münster: Waxmann.

ren erkennen?) und „Wagnis/Kreativität" (Lässt der Text ein besonderes sprachliches Wagnis erkennen? Ist er in besonderer Weise kreativ?).

Anpassungen wurden entsprechend bisheriger Erfahrungen mit der Korrektur von Texten und aufgrund testtheoretischer Gütekriterien vorgenommen. Zudem wurden die inhaltlichen Kriterien auf die konkrete Aufgabenstellung angepasst. Das Beurteilungsverfahren entspricht einem analytischen Vorgehen, bei dem verschiedene Aspekte eines Textes unabhängig voneinander nach verbal formulierten Abstufungen bewertet werden (Analytical Scoring; Weigle, 2002)[6].

## 3.2 Beurteilungskriterien

In Tabelle 1 sind die Kriterien zur Beurteilung der Texte ersichtlich. Bezüglich der Dimension *Inhalt (Auftragserfüllung und Aussagekraft)* wurden bei beiden Themen inhaltlich analoge Bereiche beurteilt: Bewertet wurde, ob die auf das Thema hinführenden Fragen „Wie sieht dein Leben als Star aus" beziehungsweise „Welche Erfahrungen machst du mit Noten" beantwortet wurden. Zusätzlich wurde bewertet, inwiefern auf die Vor- und Nachteile des gestellten Themas eingegangen, eine eigene Meinung dargelegt und wie das Gesagte mit Beispielen veranschaulicht wurde. Bei den Kriterien zur Dimension *Textaufbau und Textzusammenhang* wurde beurteilt, ob ein Text in eine sinnvolle äussere Gliederung (Einleitung, Hauptteil, Schluss) eingeteilt ist. Zusätzlich wurde bewertet, ob der Text in sich logisch zusammenhängt (Kohärenz) und ob die Sätze und Abschnitte sprachlich sinnvoll verbunden sind (Kohäsion). Bei der Dimension *Sprachrichtigkeit* wurde beurteilt, ob ein Text in Bezug auf Rechtschreibung, Grammatik, Satzbau und Satzzeichen korrekt ist. Die Dimension *Sprachangemessenheit* umfasst die Beurteilung der Wortwahl und des Satzbaus: Bewertet wurde, ob die Wortwahl angemessen und treffsicher ist und ob der Satzbau abwechslungsreich ist. Bezüglich der Dimension *Schreibstil und Kreativität* wurde beurteilt, wie gewandt sich die Schülerinnen und Schüler ausdrückten (Sprachstil), wie angemessen sie den Satzbau und die Wortwahl gestalteten sowie ob sie sprachlich und inhaltlich etwas wagten.

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

Tabelle 1: Kriterien des Beurteilungsrasters

| Dimensionen | Kriterien | Abstufungen |
|---|---|---|
| **Inhalt: Auftragserfüllung und Aussagekraft** | 1.1 Thema | 1 nur im Ansatz dargelegt<br>2 weitgehend dargelegt<br>3 eher ausführlich, detailliert dargelegt |
| | 1.2 Vor- und Nachteile | 1 nur im Ansatz dargelegt<br>2 weitgehend dargelegt<br>3 eher ausführlich, detailliert dargelegt |
| | 1.3 eigene Meinung | 1 nur im Ansatz dargelegt<br>2 weitgehend dargelegt<br>3 eher ausführlich, detailliert dargelegt |
| | 1.4 Beispiele zur Veranschaulichung | 1 keine oder unpassende Beispiele verwendet<br>2 wenige Beispiele oder wenig passende/anschauliche Beispiele verwendet<br>3 mehrere anschauliche und passende Beispiele verwendet |
| **Textaufbau und Textzusammenhang** | 2.1 Textaufbau<br>(Abschnitte – äussere Gliederung) | 1 zufällig, unüberlegt, ungegliedert oder inkonsequent<br>2 teilweise gegliedert, zwei der Komponenten (Einleitung, Hauptteil, Schluss) sind ersichtlich<br>3 Einleitung, Hauptteil und Schluss sind ersichtlich<br>4 zusätzlich zur Grundgliederung (Einleitung, Hauptteil, Schluss) in inhaltliche Sinnschritte gegliedert |
| | 2.2 logischer Zusammenhang<br>(innere Gliederung – Kohärenz) | 1 nur teilweise logisch ausgeführte Gedanken<br>2 meistens logisch ausgeführte Gedanken<br>3 logisch ausgeführte Gedanken |
| | 2.3 sprachlicher Zusammenhang<br>(innere Gliederung – Kohäsion) | 1 nur teilweise sinnvoll verbunden<br>2 meistens sinnvoll verbunden<br>3 sinnvoll verbunden, auch bei komplexeren Verbindungen |
| **Sprachrichtigkeit** | 3.1 Rechtschreibung | 1 kaum beherrscht<br>2 teilweise beherrscht<br>3 nahezu fehlerfrei |
| | 3.2 Grammatik<br>(Genus, Kasus, Tempus, Modus) | 1 kaum beherrscht<br>2 teilweise beherrscht<br>3 nahezu fehlerfrei |
| | 3.3 Satzbau | 1 kaum beherrscht<br>2 teilweise beherrscht<br>3 nahezu fehlerfrei |
| | 3.4 Satzzeichen | 1 kaum beherrscht<br>2 teilweise beherrscht<br>3 nahezu fehlerfrei |

| Dimensionen | Kriterien | Abstufungen |
|---|---|---|
| **Sprachangemessenheit** | 4.1 Satzbau | 1 sehr einfach, eintönig<br>2 etwas abwechslungsreich<br>3 abwechslungsreich, vielseitig |
| | 4.2 Wortwahl | 1 begrenzt, teilweise unangemessen<br>2 eher treffend, angemessen<br>3 treffsicher, auch bei komplexeren Begriffen |
| **Schreibstil und Kreativität** | 5.1 Schreibstil | 1 sprachlich unsicher, nicht gewandt<br>2 sprachlich wenig gewandt<br>3 sprachlich gewandt<br>4 sprachlich sehr gewandt, ausdrucksstark |
| | 5.2 Sprachliches Wagnis<br>Kreativität und Ästhetik | 1 wagt wenig, wenig kreativ<br>2 wagt etwas, etwas kreativ<br>3 wagt viel, kreativ<br>4 ausgesprochen kreativer Text, unerwartete Formulierungen |
| | 5.3 Inhaltliches Wagnis<br>Kreativität | 1 wagt wenig, wenig kreativ<br>2 wagt etwas, etwas kreativ<br>3 wagt viel, kreativ<br>4 ausgesprochen kreativer Text, unerwartete Ausführungen |

# 4 Analyse der Beurteilungskriterien

Die Qualität der Beurteilungskriterien wird anhand der üblichen Testgütekriterien der klassischen und probabilistischen Testtheorie ausgewiesen.

Die Schwierigkeit eines Beurteilungskriteriums kann durch die Lösungshäufigkeit beziehungsweise den Anteil an Schülerinnen und Schülern, welche die einzelnen Abstufungen erreicht haben, dargestellt werden. Eine Lösungshäufigkeit zwischen 75 und 100 Prozent deutet auf ein (sehr) einfaches Kriterium hin, eine Lösungshäufigkeit von 50 bis 75 Prozent auf ein eher einfaches Kriteriums, eine Lösungshäufigkeit zwischen 25 und 50 Prozent auf ein eher schwieriges Kriterium und eine Lösungshäufigkeit von weniger als 25 Prozent auf ein (sehr) schwieriges Kriterium.

Unter der Trennschärfe wird die Korrelation des Beurteilungskriteriums mit der Gesamtpunktzahl im Modul „Texte schreiben" verstanden. Der Trennschärfekoeffizient zeigt bei einem Test, inwiefern ein Bewertungskriterium Schülerinnen und Schüler mit hoher Punktzahl von Schülerinnen und Schülern mit niedriger Punktzahl trennt. Ein hoher Trennschärfekoeffizient (0.30 bis 1.00) bedeutet, dass Schülerinnen und Schüler mit einer hohen Punktzahl eine hohe Bewertung des jeweiligen Beurteilungskriteriums erreichten und solche mit einer niedrigen Punktzahl eine niedrige Bewertung. Ein niedriger Trennschärfekoeffizient (um 0) bedeutet, dass Schülerinnen und Schüler mit hohen und niedrigen Punktzahlen bei einem Kriterium gleich häufig eine hohe oder niedrige Bewertung erhielten. Ein negativer Trennschärfekoeffizient bedeutet, dass Schülerinnen und Schüler

mit einer hohen Punktzahl eine niedrige Bewertung erhielten und solche mit einer niedrigen Punktzahl eine hohe Bewertung. Dementsprechend sollte der Trennschärfekoeffizient nicht kleiner als 0.30 sein. Angewendet auf die Beurteilung von Texten zeigt die Trennschärfe, wie gut die Punktzahl bei einem Kriterium mit der Gesamtpunktzahl übereinstimmt.

Der Infit (Weighted Mean Square, MNSQ) zeigt, wie gut die einzelnen Beurteilungskriterien zum Rasch-Modell passen. Er zeigt für jedes Kriterium, wie viele unerwartete Bewertungen unter der Annahme des Rasch-Modells beobachtet wurden. Der Infit hat ein Erwartungswert von 1.0. Weichen die Infit-Werte statistisch signifikant von 1.0 ab, dann passt das Beurteilungskriterium eigentlich nicht zum Rasch-Modell. Ein zu hoher Infit-Wert weist darauf hin, dass die Trennschärfe des Beurteilungskriteriums zu niedrig ist. Ein zu tiefer Infit-Wert weist darauf hin, dass die Trennschärfe zu hoch ist. Ein guter Infit-Wert sollte in der Regel nicht kleiner als 0.7 und nicht grösser als 1.3 sein (Wright & Linacre, 1994)[7].

Der T-Wert zeigt, ob die Infit-Werte statistisch signifikant vom erwarteten Wert von 1.0 abweichen. Ein Wert grösser als 1.96 beziehungsweise kleiner -1.96 weist auf eine statistisch signifikante Abweichung hin. Die Bewertung eines Beurteilungskriteriums sollte sich allerdings nicht in erster Linie nach dem T-Wert richten, denn dieser hängt unter anderem auch von der Stichprobengrösse ab, welche beim vorliegenden Modul „Texte schreiben" mit 9953 sehr hoch ausfällt.

## 4.1 Qualität der Beurteilungskriterien

In den nachfolgenden Tabellen 2 bis 6 wird die Qualität der einzelnen Beurteilungskriterien, gegliedert in die fünf Dimensionen, dargestellt. In der ersten Spalte der Tabellen befinden sich die Bezeichnungen der Beurteilungskriterien, in der zweiten Spalte die Abstufungen zur Beurteilung der Texte anhand der Kriterien. In der dritten Spalte ist die Schwierigkeit der einzelnen Abstufungen aufgeführt. Beispielsweise wurde das Thema (siehe Tabelle 2) in 15 Prozent der Texte nur im Ansatz dargelegt; in 42 Prozent der Texte wurde das Thema weitgehend dargelegt; in 43 Prozent der Texte wurde das Thema eher ausführlich und detailliert dargelegt. Die dritte Spalte beinhaltet die Trennschärfe zur Beurteilung wie gut die Punktzahl bei einem Kriterium mit der Gesamtbeurteilung übereinstimmt. Die vierte und fünfte Spalte zeigt, wie gut die einzelnen Beurteilungskriterien zum Rasch-Modell passen.

In Tabelle 2 ist die Qualität der Beurteilungskriterien der Dimension Inhalt (Auftragserfüllung und Aussagekraft) dargestellt. Die Trennschärfe der vier Beurteilungskriterien kann als ausreichend bis gut bewertet werden. Zudem passen alle Kriterien gut zum Rasch-Modell. Die Texte verteilen sich relativ gleichmässig auf die beiden oberen Abstufungen (Stufe 2 und 3), während die unterste Stufe (1) eher seltener vertreten ist. Dies zeigt sich vor allem bei den beiden Kriterien „Vor- und Nachteile" und „Beispiele" deutlich. Nur wenige Schülerinnen und Schüler hatten die Vor- und Nachteile nur im Ansatz dargelegt oder keine beziehungsweise unpassende Beispiele zur Veranschaulichung des Gesagten verwendet.

[7] Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Tabelle 2: Qualität der Kriterien zur Dimension Inhalt (Auftragserfüllung und Aussagekraft)

| Beurteilungskriterien | Abstufungen | Schwierigkeit | Trennschärfe | Infit (MNSQ) | T-Wert |
|---|---|---|---|---|---|
| 1.1 Thema | 1<br>2<br>3 | 15%<br>42%<br>43% | 0.37 | 1.14 | 11.7 |
| 1.2 Vor- und Nachteile | 1<br>2<br>3 | 9%<br>48%<br>43% | 0.41 | 1.06 | 4.6 |
| 1.3 eigene Meinung | 1<br>2<br>3 | 28%<br>45%<br>27% | 0.39 | 1.14 | 11.9 |
| 1.4 Beispiele | 1<br>2<br>3 | 8%<br>48%<br>44% | 0.44 | 1.02 | 2.0 |

Tabelle 3 beinhaltet die Darstellung der Qualität der Beurteilungskriterien der Dimension Textaufbau und Textzusammenhang. Die Trennschärfe aller drei Kriterien ist als gut zu bewerten. Die beiden Kriterien zum logischen und sprachlichen Zusammenhang passen gut zum Raschmodell, während das Item zum Textaufbau gerade noch in einem akzeptablen Bereich liegt. Aus der Schwierigkeit der einzelnen Abstufungen bei diesem Kriterium ist ersichtlich, dass die Schülerinnen und Schüler entweder keine äussere, visuelle Gliederung gemacht haben (Stufe 1) oder aber die Dreiteilung in Einleitung, Hauptteil und Schluss gut ersichtlich ist (Stufe 3). Die Stufen 2 und 4 wurden hingegen eher selten erreicht. Bei den beiden anderen Kriterien (logischer und sprachlicher Zusammenhang) verteilen sich die Texte wiederum relativ gleichmässig auf die beiden oberen Abstufungen (2 und 3), während die unterste Stufe (1) eher seltener vertreten ist. Dies bedeutet, dass nur wenige Texte aus nur teilweise logischen Gedanken bestanden und die Sätze und Abschnitte eher selten nur teilweise sinnvoll verbunden waren.

Tabelle 3: Qualität der Kriterien zur Dimension Textaufbau und Textzusammenhang

| Beurteilungskriterien | Abstufungen | Schwierigkeit | Trennschärfe | Infit (MNSQ) | T-Wert |
|---|---|---|---|---|---|
| 2.1 Textaufbau | 1<br>2<br>3<br>4 | 29%<br>12%<br>47%<br>12% | 0.43 | 1.3 | 22.9 |
| 2.2 logischer Zusammenhang | 1<br>2<br>3 | 5%<br>52%<br>43% | 0.54 | 0.92 | -7.1 |
| 2.3 sprachlicher Zusammenhang | 1<br>2<br>3 | 8%<br>45%<br>47% | 0.58 | 0.91 | -7.4 |

In Tabelle 4 ist die Qualität der Beurteilungskriterien der Dimension Sprachrichtigkeit dargestellt. Die Trennschärfe der vier Beurteilungskriterien (Rechtschreibung, Grammatik, Satzbau und Satzzeichen) kann als gut beurteilt werden. Darüber hinaus passen die beiden Kriterien „Rechtschreibung" und „Grammatik" hervorragend zum Rasch-Modell. Bei der Rechtschreibung und bei den Satzzeichen verteilen sich die Texte relativ gleichmässig auf alle drei Abstufungen, während bei der Grammatik und beim Satzbau wiederum die unterste Stufe (1) eher seltener vertreten ist. Dieses Ergebnis bedeutet, dass relativ wenige Schülerinnen und Schüler die Grammatik und den Satzbau kaum beherrschten, während eine eher geringe Beherrschung der Rechtschreibung und der Setzung von Satzzeichen in einem etwas grösseren Ausmass vorhanden war.

Tabelle 4: Qualität der Kriterien zur Dimension Sprachrichtigkeit

| Beurteilungskriterien | Abstufungen | Schwierigkeit | Trennschärfe | Infit (MNSQ) | T-Wert |
|---|---|---|---|---|---|
| 3.1 Rechtschreibung | 1<br>2<br>3 | 25%<br>43%<br>32% | 0.53 | 0.99 | -1.3 |
| 3.2 Grammatik | 1<br>2<br>3 | 7%<br>27%<br>66% | 0.48 | 0.98 | -1.2 |
| 3.3 Satzbau | 1<br>2<br>3 | 7%<br>40%<br>53% | 0.56 | 0.92 | -6.2 |
| 3.4 Satzzeichen | 1<br>2<br>3 | 24%<br>52%<br>24% | 0.42 | 1.08 | 6.6 |

Die Tabelle 5 beinhaltet die Darstellung der Qualität der Beurteilungskriterien der Dimension Sprachangemessenheit. Die Trennschärfe der beiden Kriterien ist als gut zu bewerten und beide Kriterien passen zum Rasch-Modell. Bei beiden Kriterien verteilen sich die Texte erneut relativ gleichmässig auf die beiden oberen Abstufungen (2 und 3), während die unterste Stufe (1) eher seltener vertreten ist. Dies bedeutet, dass nur wenige Schülerinnen und Schüler einen sehr einfachen und eintönigen Satzbau sowie eine begrenzte und teilweise unangemessene Wortwahl verwendet hatten.

Tabelle 5: Qualität der Kriterien zur Dimension Sprachangemessenheit

| Beurteilungskriterien | Abstufungen | Schwierigkeit | Trennschärfe | Infit (MNSQ) | T-Wert |
|---|---|---|---|---|---|
| 4.1 Satzbau | 1<br>2<br>3 | 5%<br>46%<br>49% | 0.59 | 0.89 | -9.4 |
| 4.2 Wortwahl | 1<br>2<br>3 | 4%<br>61%<br>35% | 0.56 | 0.89 | -9.7 |

In Tabelle 6 ist die Qualität der Beurteilungskriterien zur Dimension Schreibstil und Kreativität dargestellt. Die Trennschärfe der drei Kriterien kann als gut (inhaltliches Wagnis) bis sehr gut (Schreibstil) beurteilt werden. Darüber hinaus passen die beiden Kriterien zum sprachlichen und zum inhaltlichen Wagnis sehr gut zum Rasch-Modell. Das Kriterium „Schreibstil" passt hingegen etwas weniger gut zum Rasch-Modell, die Untergrenze des MNSQ von 0.70 wird aber noch gut erreicht. Bei allen drei Kriterien trafen die unterste (Stufe 1) und die oberste Stufe (4) relativ selten auf die Texte der Schülerinnen und Schüler zu. Dies bedeutet, dass es nur wenige Texte gab, welche sprachlich nicht gewandt und wenig kreativ waren, aber gleichzeitig waren auch nur wenige Texte sprachlich sehr gewandt und ausgesprochen kreativ.

Tabelle 6: Qualität der Kriterien zur Dimension Schreibstil und zur Kreativität

| Beurteilungskriterien | Abstufungen | Schwierigkeit | Trennschärfe | Infit (MNSQ) | T-Wert |
|---|---|---|---|---|---|
| 5.1 Schreibstil | 1<br>2<br>3<br>4 | 5%<br>34%<br>59%<br>2% | 0.72 | 0.78 | -16.9 |
| 5.2 Sprachliches Wagnis<br>Kreativität und Ästhetik | 1<br>2<br>3<br>4 | 3%<br>60%<br>35%<br>2% | 0.51 | 0.93 | -5.2 |
| 5.3 Inhaltliches Wagnis<br>Kreativität | 1<br>2<br>3<br>4 | 3%<br>58%<br>38%<br>1% | 0.41 | 1.02 | 1.5 |

## 4.2 Eindimensionalität des Beurteilungsrasters

Für die Reliabilität beziehungsweise die Messgenauigkeit wird der Koeffizient „Cronbach-Alpha" berechnet. Das Cronbach-Alpha zeigt, wie stark die Beurteilungskriterien zusammenhängen.

Die Reliabilität beziehungsweise die Messgenauigkeit des gesamten Beurteilungsrasters erreicht ein Cronbach-Alpha von 0.78. Dies weist darauf hin, dass Schülerinnen und Schüler mit einer hohen Schreibkompetenz bei allen Kriterien eine höhere Bewertung erhielten als Schülerinnen und Schüler mit einer geringeren Schreibkompetenz. Die Beurteilungskriterien sind somit ziemlich konsistent angewendet worden und eignen sich relativ gut, um zuverlässig zwischen guten und weniger guten Texten zu differenzieren und um die Schreibkompetenz von Schülerinnen und Schülern zu bestimmen. Wie zuverlässig die Korrektur der Texte durchgeführt wurde, ist im folgenden Kapitel 5 detailliert beschrieben.

# 5 Beurteilungszuverlässigkeit

## 5.1 Zusammensetzung des Korrekturteams

Die Texte der Schülerinnen und Schüler wurden von sieben Korrektorinnen und Korrektoren (Rater) bewertet. Rater A studierte Pädagogik und Deutsche Sprach- und Literaturwissenschaft. Rater B studierte Deutsche Sprach- und Literaturwissenschaft. Rater C studierte Deutsche Sprach- und Literaturwissenschaft. Rater D ist eine ausgebildete und erfahrende Sekundarlehrperson phil I. Rater E studierte Deutsche Sprach- und Literaturwissenschaft und schliesst im Mai 2013 die Ausbildung zum Gymnasiallehrer ab. Rater G studierte Deutsche Sprach- und Literaturwissenschaft. Rater F ist ausgebildete Sekundarlehrperson phil I und unterrichtet seit zwei Jahren auf dieser Stufe. Die Zuordnung der Texte zu den Ratern erfolgte klassenweise nach dem Zufallsprinzip.

## 5.2 Verfahren der Doppelkorrekturen

Es besteht die Gefahr, dass sich bei verschiedenen Ratern über die Zeit hinweg ein unterschiedliches mentales Modell einzelner Beurteilungskriterien entwickelt. Daher ist es wichtig in regelmässigen Abständen das gemeinsame Verständnis der Beurteilungskriterien zu überprüfen (Hoyt, 2000)[8]. Dies wurde dadurch gewährleistet, indem täglich zehn Texte von allen anwesenden Ratern beurteilt wurden (Doppelkorrekturen). Das Ziel war dabei, eine möglichst hohe Übereinstimmung der Beurteilung der Texte zwischen den Ratern zu erreichen. Insgesamt wurden mit diesem Verfahren rund 360 Texte von mehreren Ratern beurteilt. Die Abweichungen in der Beurteilung wurden fortlaufend an die Rater zurückgemeldet und besprochen.

## 5.3 Übereinstimmung der Bewertung

In Tabelle 7 sind die Übereinstimmungen beziehungsweise die Abweichungen zwischen den Ratern bei den Doppelkorrekturen dargestellt. Die Prozentzahl der Übereinstimmung beziehungsweise der Abweichung kommt über einen Vergleich von allen sieben Ratern untereinander zustande (21 mögliche Kombinationen). Die Anzahl an Vergleichen pro Kriterium ist in der zweiten Spalte dargestellt. Die Spalte „Übereinstimmung" beinhaltet den Anteil der vollständigen Übereinstimmung zweier Rater, die darauf folgenden Spalten enthalten dementsprechend den Anteil an Abweichungen um einen, zwei oder drei Punkte bei den Abstufungen pro Kriterium.

[8] Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64–86.

Tabelle 7: Übereinstimmung und Abweichungen in den Bewertungen

| Beurteilungskriterien | N | Übereinstimmung | Abweichung 1 Punkt | Abweichung 2 Punkte | Abweichung 3 Punkte |
|---|---|---|---|---|---|
| 1.1 Thema | 4733 | 56% | 39% | 5% | – |
| 1.2 Vor- und Nachteile | 4783 | 58% | 40% | 2% | – |
| 1.3 eigene Meinung | 4778 | 50% | 44% | 6% | – |
| 1.4 Beispiele zur Veranschaulichung | 4778 | 56% | 42% | 2% | – |
| 2.1 Textaufbau | 4783 | 73% | 21% | 5% | 1% |
| 2.2 logischer Zusammenhang | 4777 | 52% | 46% | 2% | – |
| 2.3 sprachlicher Zusammenhang | 4777 | 49% | 49% | 2% | – |
| 3.1 Rechtschreibung | 4778 | 60% | 38% | 2% | – |
| 3.2 Grammatik | 4783 | 68% | 31% | 1% | – |
| 3.3 Satzbau | 4780 | 58% | 41% | 1% | – |
| 3.4 Satzzeichen | 4778 | 55% | 42% | 3% | – |
| 4.1 Satzbau | 4777 | 55% | 43% | 2% | – |
| 4.2 Wortwahl | 4777 | 58% | 41% | 1% | – |
| 5.1 Schreibstil | 4783 | 67% | 32% | 1% | 0% |
| 5.2 Sprachliches Wagnis | 4783 | 61% | 38% | 1% | 0% |
| 5.3 Inhaltliches Wagnis | 4783 | 61% | 38% | 1% | 0% |

Wie Tabelle 7 zeigt, ist die prozentuale Übereinstimmung beim sprachlichen Zusammenhang (Kriterium 2.3) mit 49 Prozent am geringsten. Dies bedeutet, dass die Rater sich in etwa der Hälfte aller Vergleiche uneinig waren und einen Punkt mehr oder weniger vergeben haben als der andere Rater. Die höchste prozentuale Übereinstimmung liegt mit 73 Prozent beim Kriterium Textaufbau (2.1). Die angestrebte möglichst hohe Übereinstimmung zwischen den Personen des Korrekturteams wurde somit insgesamt gesehen ansatzweise erreicht. Bei den drei Kriterien „Textaufbau" (2.1), „Grammatik" (3.2) und „Schreibstil" (5.1) wurde die angestrebte Übereinstimmung gut erreicht. Bei den drei Kriterien „eigene Meinung" (1.3), „logischer Zusammenhang" (2.2) und „sprachlicher Zusammenhang" (2.3) ist die Übereinstimmung hingegen noch als verbesserungswürdig zu bewerten.

Nicht bestimmt werden kann die Korrektur-Strenge der sieben Rater des Korrekturteams bei der Darstellung der Übereinstimmung zwischen den Ratern. Mit der Anwendung der Item-Response-Theorie ist es möglich, die Korrektur-Strenge der beurteilenden Personen ins Testmodell einzubeziehen und bei der Berechnung der Ergebnisse zu berücksichtigen. Ein solches Vorgehen wird auch

als „Multi-Facetten-Modell" bezeichnet[9]. Die Strenge der Rater wird dabei auf einer logarithmischen Skala abgebildet und als „Logit" bezeichnet. Ein positiver Logit entspricht einer hohen Strenge, ein negativer Logit einer geringen Strenge. Wie Tabelle 8 zeigt, variiert die Strenge der Rater zwischen Rater C mit einem Logit von -0.296 und Rater F mit einem Logit von 0.295. Rater C hat somit am wenigsten streng korrigiert, während Rater F die Texte der Schülerinnen und Schüler am strengsten beurteilt hat. Die Spannweite der Beurteilung beträgt somit 0.59 Logits. Angesichts der Standardabweichung von 0.77 Logits in der Populationsverteilung entspricht dies einer mittelstarken Differenz. Die Bewertungen der Rater sind alle hinreichend modellkonform und passen somit zum Rasch-Modell. Dies bedeutet, dass die Bewertungen der Texte primär durch die Unterschiede in der Qualität der Texte im Bereich „Texte schreiben" erklärt werden können.

Tabelle 8: Strenge und Modellkonformität der Korrekturen

| Rater | Strenge (Logit) | Schätzfehler | Infit (MNSQ) | T-Wert |
|---|---|---|---|---|
| Rater A | -0.238 | 0.007 | 1.04 | 1.1 |
| Rater B | -0.146 | 0.007 | 1.02 | 0.6 |
| Rater C | -0.296 | 0.007 | 0.99 | -0.2 |
| Rater D | 0.227 | 0.007 | 1.12 | 3.4 |
| Rater E | 0.113 | 0.007 | 1.00 | -0.0 |
| Rater F | 0.295 | 0.007 | 0.99 | -0.2 |
| Rater G | 0.045 | 0.018 | 0.92 | -2.2 |

## 5.4 Stabilität der Bewertung

Die „Korrektur-Strenge" nimmt über die Zeit zu, bei ca. 20% der Rater sogar in einem sehr hohen Ausmass. Daher ist es wichtig, dass ein konstante Kontrolle der Rating-Stabilität stattfindet (Congdon & McQueen, 2000)[10]. Um die Rating-Leistung über die Zeit verfolgen zu können, wurde die Korrekturphase in Zeiteinheiten unterteilt. Bei der Methode der „baseline comparison" (Myford & Wolfe, 2009)[11] wird das Rating einer Person zum Zeitpunkt x mit dem Rating zu einem bestimmten Referenzpunkt verglichen.

Im Rahmen der Korrekturen des Moduls „Texte schreiben" wurden nach etwa zwei Wochen Korrekturzeit 15 Aufsätze von allen sieben Ratern korrigiert (Referenzpunkt). Jeweils drei dieser Aufsätze wurden in den darauf folgenden fünf Korrekturwochen in die Doppelkorrekturen eingestreut und somit erneut bewertet. Die Punktzahlen der beiden Bewertungen wurden miteinander verglichen (Vergleich Referenzpunkt mit erneuter Korrektur). Im Gegensatz zu bisherigen empirischen Befunden konnte keine kontinuierliche Zu- oder Abnahme der Korrektur-Strenge über die Zeit beobachtet werden.

---

[9] Linacre, J. M. 1994. *Many-Facet Rasch Measurement.* Chicago, IL: MESA Press (original work published in 1989).

[10] Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in Large-Scale Assessment Programs. *Journal of Educational Measurement, 37*, 163–178.

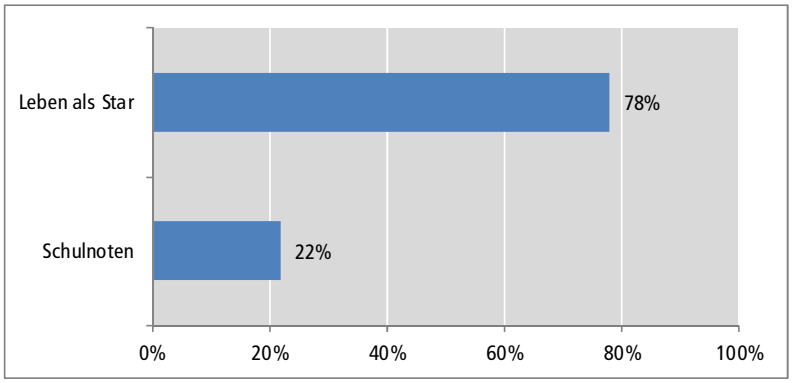
[11] Myford, C.M & Wolfe, E.W (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*, 371–389.

## 5.5 Auswahl und Schwierigkeit der Themen

Abbildung 4 zeigt, welches Thema von wie vielen Schülerinnen und Schülern gewählt wurde. Das Thema „das Leben als Star" war bei den Schülerinnen und Schüler beliebter, 78 Prozent wählten dieses Thema für ihren Aufsatz. Nur knapp ein Viertel der Schülerinnen und Schüler (22%) wählten das Thema „Muss es in der Schule Noten geben?".

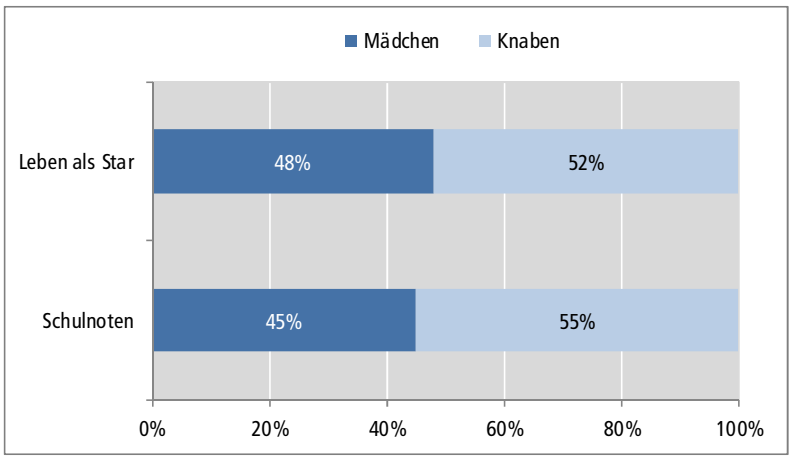
Dass der grösste Teil der Schülerinnen und Schüler das Thema „das Leben als Star" wählten, liegt vermutlich daran, dass dieses Thema auf den ersten Blick einfacher erscheint und etwas mehr Spielraum für die Fantasie der Schülerinnen und Schüler offen lässt.

Abbildung 4: Auswahl der Themen




In der Abbildung 5 ist die Wahlpräferenz nach Geschlecht dargestellt. Es zeigt sich, dass beide Themen in etwa gleich oft von Mädchen oder von Knaben gewählt wurden.

Abbildung 5: Auswahl der Themen nach Geschlecht




Mit Hilfe der Raschskalierung ist es möglich, die Schwierigkeit der beiden Themen unabhängig von den Fähigkeiten der Schülerinnen und Schüler sowie unabhängig von der Strenge der Rater zu schätzen. In Tabelle 9 ist das Ergebnis der Skalierung dargestellt. Auch die Schwierigkeit der Themen kann im Rasch-Modell geschätzt werden und auf einer logarithmischen Skala als „Logit" abgebildet werden. Je positiver der Logit ausfällt, desto höher ist die Schwierigkeit. Aus den Logits

(Schwierigkeit) in Tabelle 9 ist ersichtlich, dass beide Themen in etwa gleich streng beurteilt wurden. Die Wahl des Themas wurde deshalb in den weiteren Analysen nicht berücksichtigt. Die Vermutung, dass das Thema „das Leben als Star" für die Schülerinnen und Schüler einfacher sein könnte und daher öfter gewählt wurde, hat sich somit nicht bestätigt.

Tabelle 9: Schwierigkeit und Modellkonformität der beiden Themen

| Thema | Schwierigkeit (Logit) | Schätzfehler | MNSQ | T-Wert |
|---|---|---|---|---|
| Leben als Star | 0.065 | 0.004 | 0.98 | -0.9 |
| Schulnoten | -0.065 | 0.004 | 1.02 | 0.6 |

# 6 Testergebnisse

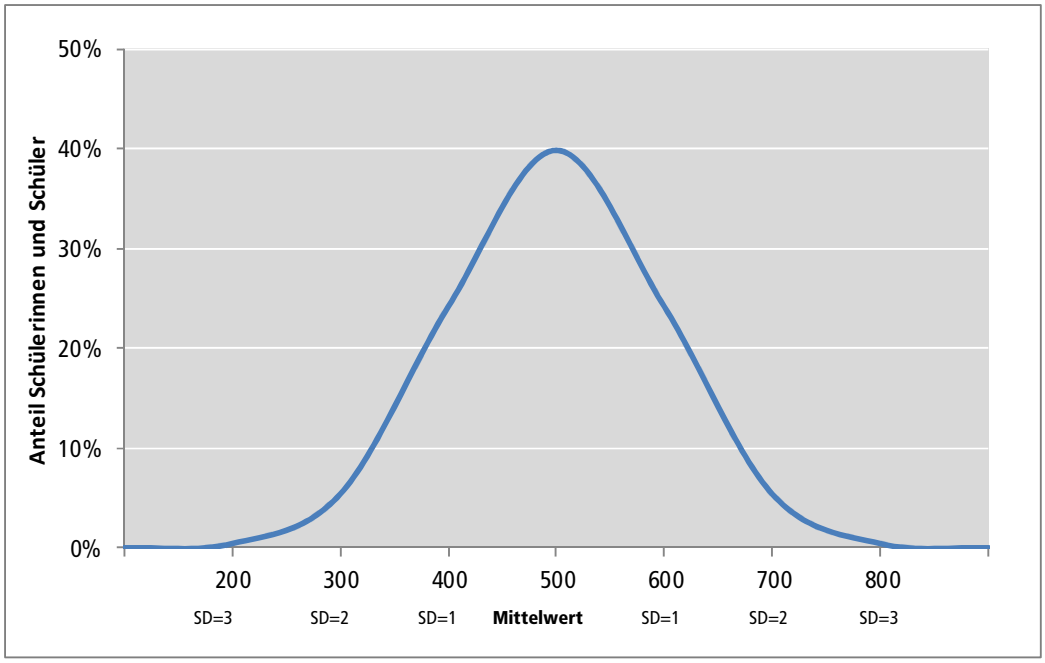
Dass derselbe Text trotz vorgegebener, standardisierter Beurteilungskriterien, einer Schulungsphase der Rater und regelmässiger Doppelkorrekturen von mehreren Personen immer gleich beurteilt wird, ist aufgrund des Interpretationsspielraums bei offen gestellten Aufgaben nicht zu erwarten. Bei der Beurteilung eines Textes bestimmen vier Faktoren das Testergebnis: a) Die Schreibkompetenz der Schülerin oder des Schülers; b) die Schwierigkeit des Beurteilungskriteriums; c) die Korrektur-Strenge der Rater und d) das Thema. Die Fähigkeiten der Schülerinnen und Schüler, die Beurteilungsstrenge der Rater und das Thema werden als Facetten der Beurteilungssituation aufgefasst und bei der Berechnung der Ergebnisse berücksichtigt. Alle Facetten werden unabhängig voneinander geschätzt und auf derselben logarithmischen Skala (Logit) abgebildet.

Wie weiter oben in Tabelle 8 dargestellt, liegt die maximale Differenz der Strenge der Beurteilung der Texte bei 0.59 Logits, was einer mittelstarken Differenz entspricht. Dieser Wert entspricht der Differenz zwischen dem strengsten Rater F und dem mildesten Rater C. Die unterschiedliche Strenge bei der Beurteilung wurde deshalb bei der Berechnung der Ergebnisse der Schülerinnen und Schüler berücksichtigt. Das Thema wurde hingegen in den Auswertungen nicht berücksichtigt, da die Schwierigkeit beider Themen sehr ähnlich ausgefallen ist (siehe Tabelle 9) und zudem nicht ausgeschlossen werden kann, dass ein Thema nur deshalb minim schwieriger ausgefallen ist, weil es in der Tendenz eher von schwächeren Schülerinnen und Schülern gewählt wurde.

## 6.1 Verteilung der Ergebnisse auf der Stellwerkskala

Um die Ergebnisse innerhalb des Moduls „Texte schreiben" vergleichbar zu machen, wurden die Fähigkeiten der Schülerinnen und Schüler (Logits) in eine standardisierte Normalverteilung transformiert, welche analog zur Stellwerk-Skala, einen Mittelwert von 500 und eine Standardabweichung von 100 Punkten aufweist (siehe Abbildung 6). 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 Punktzahl auf der Stellwerk-Skala zeigt den Schülerinnen und Schülern, wie gut sie innerhalb der Vergleichsgruppe der 9953 Schülerinnen und Schüler aus der 8. Klasse im Kanton Zürich – die den Text geschrieben haben – abgeschnitten haben.

Abbildung 6: Verteilung der Punkte auf der Stellwerkskala

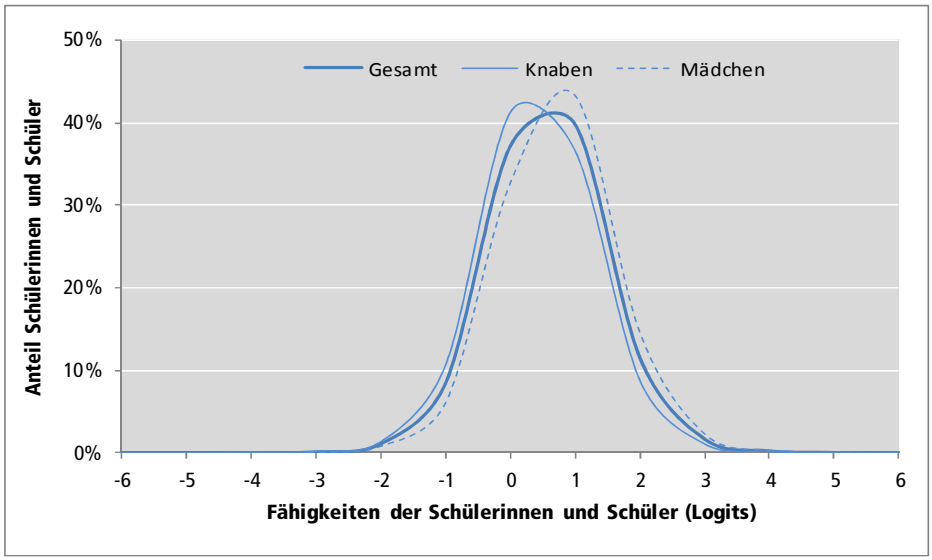



## 6.2 Verteilung der Fähigkeiten der Schülerinnen und Schüler nach Geschlecht

Die Abbildung 7 zeigt die Verteilung der Fähigkeiten der Schülerinnen und Schüler im Bereich „Texte schreiben" als Logits nach der Korrektur der unterschiedlichen Strenge der Rater. Je höher der Logit eines Schülers respektive einer Schülerin ausfällt, desto höher ist die Fähigkeit ausgeprägt. Die Fähigkeiten der Schülerinnen und Schüler sind zwischen -5 Logits und +6 Logits näherungsweise normalverteilt (siehe Abbildung 7). Die mittlere Fähigkeit liegt bei 0.58 Logits. Die schlechtesten 10 Prozent der Schülerinnen und Schüler weisen Fähigkeiten von unter -0.50 Logits auf, während die besten 10 Prozent der Schülerinnen und Schüler eine Fähigkeit von über 1.63 Logits aufweisen.

Wird die mittlere Fähigkeit der Schülerinnen und Schüler getrennt nach Geschlecht betrachtet, so weisen die Mädchen im Bereich „Texte schreiben" im Durchschnitt eine Fähigkeit von 0.73 Logits, während die durchschnittlichen Fähigkeiten der Knaben bei 0.45 Logits liegen. Dies bedeutet, dass die Mädchen im Bereich „Texte schreiben" höhere Fähigkeiten aufweisen im Vergleich zu den Knaben. Die Verteilung der Mädchen liegt dementsprechend auch etwas weiter nach rechts verschoben im Vergleich zur Verteilung der Knaben (siehe Abbildung 7). Diese Differenz in der Fähigkeit von 0.28 Logits ist statistisch signifikant (Effektstärke d = .32).

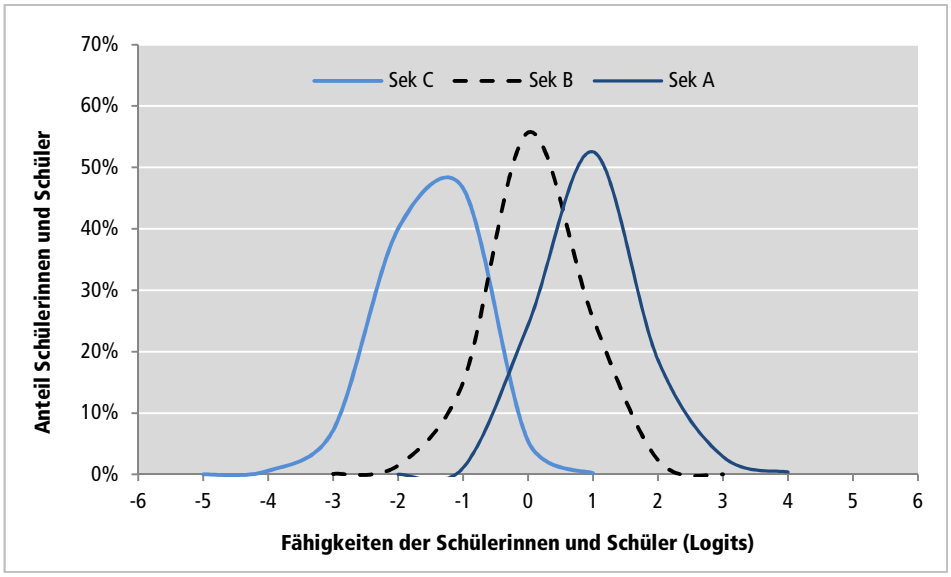
Abbildung 7[12]: Fähigkeiten nach Geschlecht im Bereich „Texte schreiben"




## 6.3 Verteilung der Fähigkeiten der Schülerinnen und Schüler nach Schulstufe

In Abbildung 8 ist die Verteilung der Fähigkeiten der Schülerinnen und Schüler der Sek A, der Sek B und der Sek C im Bereich „Texte schreiben" dargestellt.

Abbildung 8: Fähigkeiten der Schülerinnen und Schüler (Logits) im Modul „Texte schreiben"

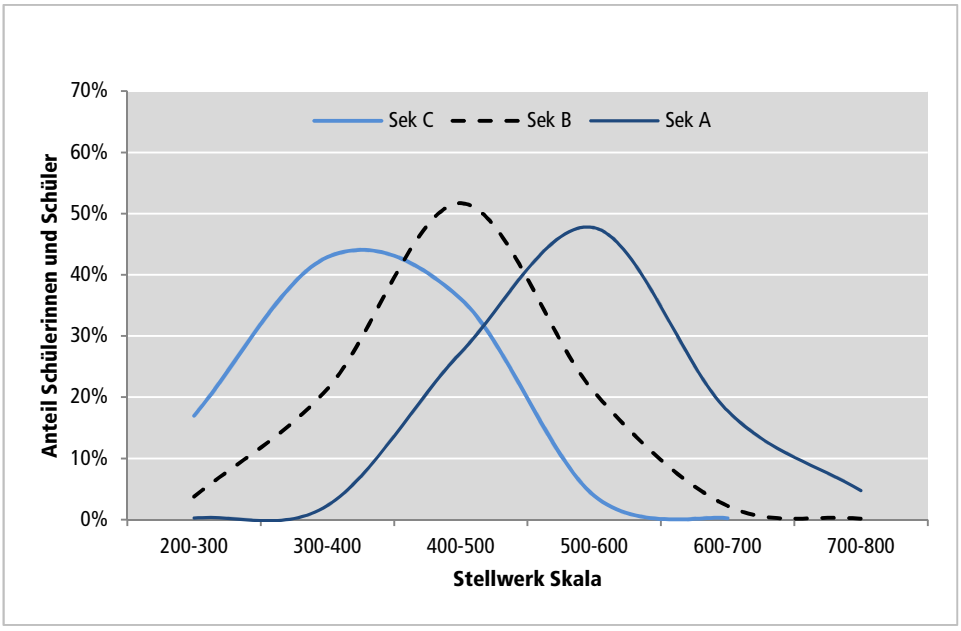



[12] *Anmerkung zur Abbildung 7:* Die Darstellung beruht auf einer Kategorisierung der Fähigkeiten der Schülerinnen und Schüler und dient der Illustration der Ergebnisse im Text. Durch die Kategorisierung können die Werte in der Grafik leicht von den exakten Werten im Text abweichen. Dasselbe trifft auch auf die Abbildungen 8 und 9 zu.

Wird die durchschnittliche Fähigkeit der Schülerinnen und Schüler getrennt nach Schulstufe betrachtet, so weisen die Schülerinnen und Schüler der Sek C im Bereich „Texte schreiben" im Durchschnitt eine Fähigkeit von -0.51 Logits auf. Die durchschnittliche Fähigkeit der Sek B Schülerinnen und Schüler liegt bei 0.13 Logits. Wie erwartet weisen die Schülerinnen und Schüler der Sek B höhere Kompetenzen im Bereich „Texte schreiben" auf als die Schülerinnen und Schüler der Sek C. Dieser Unterschied von 0.64 Logits ist statistisch signifikant und bedeutsam (Effektstärke d = .90). Wie in Abbildung 7 ersichtlich liegt die durchschnittliche Fähigkeit im Bereich „Texte schreiben" der Schülerinnen und Schüler der Sek A mit 1.00 Logits über den Fähigkeiten der Schülerinnen und Schüler der Sek B. Diese Differenz von 0.87 Logits ist wiederum statistisch signifikant und bedeutsam (Effektstärke $d$ = 1.2).

Als Ergänzung dient in Abbildung 8 die Stellwerkskala als Grundlage anstelle der Logits aus Abbildung 7. Deutlich wird mit dieser Darstellung, dass die Mehrheit der Schülerinnen und Schüler der Sek C zwischen 300 und 400 Punkte erreicht hat, während die Mehrheit der Schülerinnen und Schüler der Sek B eine Punktzahl zwischen 400 und 500 erzielte. Der Grossteil der Schülerinnen und Schüler der Sek A liegt damit über dem Mittelwert der Stellwerkskala zwischen 500 und 600 Punkten.

Abbildung 9: Fähigkeiten der Schülerinnen und Schüler auf der Stellwerkskala im Bereich „Texte schreiben"




Interessant ist, dass 21 Prozent der Schülerinnen und Schüler der Sek C Texte schreiben, deren Beurteilung über dem Mittelwert der Sek B liegt. Rund 10 Prozent der Schülerinnen und Schüler der Sek B schreiben Texte, die über dem Mittelwert der Sek A liegen. Dieses Ergebnis verdeutlicht, wie sinnvoll eine schultypenunabhängige Beurteilung sein kann.

# 7 Zusammenhang Teilbereiche im Fach Deutsch

Der Vergleich des Moduls „Texte schreiben" mit den vier computergestützten Tests „Hören und Verstehen", „Lesen und Verstehen", „Schreibfertigkeiten" sowie „Sprachreflexion und Rechtschreibung" ermöglicht ein vertieftes Verständnis darüber, wie die produktiven und die reproduktiven Fähigkeiten im Fach Deutsch zusammenhängen.

Die Ergebnisse der computerbasierten Stellwerk-Tests werden – wie das Modul „Texte schreiben" – auf einer normierten Skala von 200 bis 800 Punkten abgebildet. Die in Tabelle 10 dargestellten Mittelwerte und Standardabweichungen[13] zeigen, dass der Mittelwert bei den vier computergestützten Stellwerk-Tests tiefer liegt als beim Modul „Texte schreiben". Dies bedeutet, dass die Schülerinnen und Schüler aus dem Kanton Zürich im Modul „Texte schreiben" im Durchschnitt eine höhere Punktzahl erzielen als in den computerbasierten Tests. Diese Differenz basiert jedoch nicht auf höheren Fähigkeiten der Schülerinnen und Schüler im Bereich „Texte schreiben" sondern auf der im Kanton St. Gallen durchgeführten Normierung der computerbasierten Stellwerk-Tests.

Tabelle 10: Mittelwerte in den Teilbereichen im Fach Deutsch

| Bereich | M | SD |
|---|---|---|
| Hören und Verstehen | 473.52 | 142.37 |
| Lesen und Verstehen | 478.76 | 147.21 |
| Schreibfertigkeiten | 488.12 | 135.81 |
| Sprachreflexion / Rechtschreibung | 468.30 | 138.32 |
| Texte schreiben | 500.52 | 98.00 |
| Gesamtwert Deutsch | 474.04 | 128.00 |

*Anmerkung*: M = Mittelwert, SD = Standardabweichung

Weitere Informationen zur Validität des Gesamtwerts im Fach Deutsch liefern die Korrelationen zwischen den einzelnen Bereichen sowie die Korrelationen der Bereiche mit dem Gesamtwert. Alle entsprechenden Korrelationen sind in Tabelle 11 dargestellt.

Der Vergleich der Korrelationen zwischen den einzelnen Bereichen ergibt, dass das Modul „Texte schreiben" mit allen Bereichen im Fach Deutsch gleich hoch korreliert. Die Abweichungen zwischen den Korrelationen sind unbedeutend. Die Betrachtung der Korrelationen der Teilbereiche mit dem Gesamtwert im Fach Deutsch zeigt, dass die computergestützt erfassten Bereiche sehr hoch mit dem Gesamtwert korrelieren ($r$ zwischen 0.80 und 0.85), während das Modul „Texte schreiben" nur mit $r = 0.64$ mit dem Gesamtwert korreliert. Dieses Ergebnis ist so zu interpretieren, dass das Verfassen von Texten beziehungsweise die produktiven Fähigkeiten sich deutlich vom computerbasiert erfassten Wissen unterscheiden. Die Ergänzung der computerbasierten Stellwerk-Tests mit dem Modul „Texte schreiben" ist somit wichtig, da das Modul positiv zu einem validen Stellwerk-Test im Fach Deutsch beiträgt.

[13] Die Standardabweichung illustriert die Streubreite der Werte eines Merkmals rund um dessen Mittelwert.

Tabelle 11: Korrelationen zwischen den Teilbereichen im Fach Deutsch

| | Hören und Verstehen | Lesen und Verstehen | Schreibfertigkeiten | Sprachreflexion / Rechtschreibung | Texte schreiben | Gesamtwert Deutsch |
|---|---|---|---|---|---|---|
| Hören und Verstehen | – | 0.58 | 0.59 | 0.53 | 0.49 | 0.80 |
| Lesen und Verstehen | 0.58 | – | 0.60 | 0.57 | 0.51 | 0.83 |
| Schreibfertigkeiten | 0.59 | 0.60 | – | 0.61 | 0.56 | 0.85 |
| Sprachreflexion / Rechtschreibung | 0.53 | 0.57 | 0.61 | – | 0.56 | 0.82 |
| Texte schreiben | 0.49 | 0.51 | 0.56 | 0.56 | – | 0.64 |
| Gesamtwert Deutsch | 0.80 | 0.83 | 0.85 | 0.82 | 0.64 | – |

# 8 Beispieltexte zu den einzelnen Kompetenzniveaus

Um die Kompetenzen der einzelnen Schüler standardisiert beschreiben zu können, wurden die Punktzahlen auf der Stellwerkskala zu Kompetenzniveaus zusammengefasst (siehe Tabelle 12). Die Schreibkompetenzen, die eine Schülerin oder ein Schüler innerhalb eines bestimmten Kompetenzniveaus aufweist, wurden zum einen bezüglich der drei Teilgebiete „Inhalt", „Textaufbau" und „Sprache" beschrieben und zum anderen mit Beispieltexten illustriert. Die Beschreibung der Schreibkompetenzen auf den verschiedenen Niveaus bezieht sich jeweils auf einen durchschnittlichen Text des jeweiligen Niveaus. Bei konkreten Einzelleistungen von Schülerinnen und Schülern kann es vorkommen, dass die Kompetenzen innerhalb einer Gesamtpunktzahl je nach Teilbereich variieren. Beispielsweise könnte ein Text des Intervalls 501 bis 600 Punkte im Teilbereich „Inhalt" tiefer (z.B. im Kompetenzniveau von 401 bis 500 Punkten) und dafür im Teilbereich „Sprache" etwas höher (z.B. im Kompetenzniveau von 601 bis 700 Punkten) liegen. Innerhalb eines bestimmten Kompetenzniveaus besteht somit eine recht grosse Bandbreite an Texten, die sich innerhalb der einzelnen Teilbereiche unterscheiden können. In der folgenden Zusammenstellung wird jedes Kompetenzniveau mit einem Beispieltexten veranschaulicht. Weitere Textbeispiele befinden sich unter www.ibe.uzh.ch/projekte/texteschreibenzuerich12 oder im Praxisteil des vorliegenden Berichts.

Tabelle 12: Kompetenzbeschreibungen nach Punkteintervallen

| Punkte-intervall | Kompetenzbeschreibungen | | |
|---|---|---|---|
| | Inhalt | Textaufbau | Sprache |
| 200 bis 300 Punkte | Die Texte haben wenig Aussagekraft und die in der Aufgabenstellung verlangten Inhalte sind zwar erkennbar, werden jedoch nur im Ansatz dargelegt. Beispiele zur Veranschaulichung sind selten vorhanden, oft unpassend oder veranschaulichen das Gesagte wenig. | Der Textaufbau ist zufällig, unüberlegt, ungegliedert oder inkonsequent. Die Sätze und Abschnitte sind nur teilweise logisch und sinnvoll miteinander verbunden. | Die Rechtschreibung und die Zeichensetzung werden nicht beachtet. Grammatikalische und syntaktische Strukturen sind sehr einfach oder sehr fehlerhaft. Die sprachliche Ausdrucksweise wird als unsicher beurteilt. |
| 301 bis 400 Punkte | Die Texte erfüllen die Aufgabenstellung im Ansatz. Vor- und Nachteile werden weitgehend dargelegt. Zur Veranschaulichung werden wenige Beispiele verwendet oder aber Beispiele, die nicht genau passen beziehungsweise wenig anschaulich sind. | Die Texte sind eher zufällig gegliedert, bestehen aber aus meistens logisch ausgeführten Gedanken. Die Sätze und Abschnitte sind meistens sinnvoll miteinander verbunden. | Die Rechtschreibung und die Zeichensetzung werden kaum beachtet. Der Satzbau und die Grammatik werden teilweise beherrscht. Die Wortwahl ist meistens treffend und angemessen. Insgesamt wird die sprachliche Ausdrucksweise jedoch als wenig gewandt beurteilt. |
| 401 bis 500 Punkte | Die Texte erfüllen die Aufgabenstellung weitgehend. Mehrere Vor- und Nachteile zum Thema werden vorgestellt und mit einzelnen Beispielen illustriert. Eine eigene Meinung ist weitgehend vorhanden. | Die Texte sind teilweise gegliedert und bestehen meistens aus logisch ausgeführten Gedanken. Zumeist werden auch die Textelemente sinnvoll miteinander verbunden. | Die Rechtschreibung, Zeichensetzung und der Satzbau werden teilweise beherrscht. Die Grammatik ist nahezu fehlerfrei. Die Wortwahl ist treffend und angemessen; der sprachliche Ausdruck ist recht gewandt. |
| 501 bis 600 Punkte | Die Texte erfüllen die Aufgabenstellung und erläutern das Thema mit Vor- und Nachteilen recht ausführlich und detailliert. Eine eigene Meinung wird vertreten. Die Texte sind recht kreativ und veranschaulichen das Gesagte mit mehreren passenden Beispielen. | In den Texten ist eine Grundgliederung in Einleitung, Hauptteil und Schluss erkennbar. Die Texte bestehen aus logisch ausgeführten Gedanken und die Textelemente werden sinnvoll miteinander verbunden. | Die Rechtschreibung, die Setzung der Satzzeichen und die Grammatik sind fast fehlerfrei. Abwechslungsreich, vielseitig und fast immer korrekt ist auch der Satzbau. Die Wortwahl ist treffsicher und die sprachliche Ausdrucksweise gewandt und kreativ. |
| 601 bis 700 Punkte | Die Texte erfüllen die Aufgabenstellung und erläutern das Thema mit Vor- und Nachteilen ausführlich und detailliert. Auch die eigene Meinung wird ausführlich und begründet vertreten. Das Gesagte wird gut veranschaulicht. Die Texte sind kreativ und Neues wird gewagt. | Die Texte sind in sinnvolle Abschnitte gegliedert, ein roter Faden ist erkennbar. Die Gedankenführung ist logisch und die Textelemente werden auch bei komplexeren Formulierungen sinnvoll miteinander verbunden. | Die Rechtschreibung, die Grammatik, der Satzbau und auch die Setzung der Satzzeichen sind nahezu fehlerfrei. Die Wortwahl ist auch bei komplexeren Begriffen treffsicher und angemessen. Die Texte enthalten komplexe, abwechslungsreiche Formulierungen und sind sprachlich gewandt sowie kreativ. |
| 701 bis 800 Punkte | Die Texte gehen ausführlich, detailliert und elaboriert auf die Aufgabenstellung ein. Begründungen und Stellungnahmen sind einleuchtend und ausführlich. Inhaltlich enthalten die Texte unerwartete Ausführungen und sind ausgesprochen kreativ. | Die Texte haben einen roten Faden, folgen einem logischen, klaren Aufbau und sind äusserlich und innerlich sinnvoll gegliedert. Die Textelemente werden auch bei komplexeren Verbindungen logisch und sinnvoll miteinander verbunden. | Die Texte sind auch bei komplexeren Formulierungen nahezu fehlerfrei. Die sprachliche Ausdrucksweise ist sehr gewandt und ausdrucksstark. Die Texte zeichnen sich durch ansprechende, abwechslungsreiche und kreative Sprachstrukturen sowie unerwartete Formulierungen aus. |

Textbeispiel 1: 200 bis 300 Punkte

leben als Sportler star. Ich spiele fussball. Die Beste
von welt und Hapsch. Ich verdiene gut viel geld
Ich lache immer und ich bin Glachlich
Ich denke das ich die beste Glachliche
Mensch von der welt ich habe alles was
ich will. Jeden tag neuen kleider. teure
kleider manke kleider. Ich habe viele
fance viel leute kannen mich und gerne
wie ich ~~spiele~~. fussball spiele. Ich habe
immer 2 sekuryti dabei sie beide passens
mich auf und die leute wollen immer
mit mir foto machen und unter schrift
machen. wenn ich eine spiel spiele wenn
ich den ball habe die leute freut weil
ich schasse ein toor. Ich mache viele
verbung über schuhe und ban viele sachen.
Ich bin wie ein konig für die leute wo die
mich sehr gerne haben. Ich muss immer
trainieren viel, oft und sehr streng und
schwirig. aber ich mache etwas wo schwer
ist darum habe ich so ein gutes leben
Ich kann aber gehen wenn ich spielen muss
manchmall geht mann auch in die andern
land. Ich habe eine schöne familie und
reich schönes auto. —

Textbeispiel 2: 301 bis 400 Punkte

Ich würde eine Eiskunstlaufstar sein weil ich das Eiskunstlaufen Liebe. Aber mein Leben als Star würde Stressig sein, man wäre immer unter druck. Man würde sich immer fragen „war das jetzt gut so?". Es kann auch gelassen sein, man reist herum und zeigt was man kann. Die Meisten wollen Modelstar werden aber Sie wissen nicht wie anstrengent das sein kann, sie müssen schauen das sie nicht zu dik sind, das sie schön aussehen, usw. Es ist so wenn man ein Leben als Star hat. So wie bei jedem Leben hat es vor-und Nachteile, das hat auch das Leben als Star. Die Vorteile im Leben als Star sind: Das man berühmt ist, das reisen, gut behandelt wird, usw. Aber es giebt auch Nachteile und zwar: immer unter druck, ob man gut oder schlecht war, das gerade Stehen, nett und freundlich sein auch wenn man dem Anderen den Hals umdrehen will und man sollte immer die Ruhe bewaren. Über haupt zählt für mich nur eins die vorteile aber auch das ich mich wohl fühle bei dem was ich mache und nicht die Nachteile oder ich werd gut belohnt aber es scheisst mich an es zu tun. Ich meine ein Star zu sein Träumt jeder aber niemand nimmt es ernst wenn jemand sagt das ist nicht so toll wie ihr euch das so vorstellt. Doch jeder sagt auch wenn es so ist ich will trotz dem ein Star sein. Egal wie fest man es sich wünscht es wird nur sehr selten war. Darum Sollte man sich nicht zu viele Hoffnungen machen, es kann auch alles anderst kommen.

Textbeispiel 3: 401 bis 500 Punkte

# Mein Leben als Fussballprofi

Ich wäre am liebsten Fussballprofi, denn Fussball ist nicht nur ein Hobby, es ist ein Teil von mir. Es ist schwierig sich das Leben eines Stars vorzustellen, ~~aber~~ eins ist sicher es hätte Vor- und Nachteile. Ich würde bei meinem Lieblingsverein Real Madrid in Spanien spielen. Ich wäre Vermögend und könnte ein schönes Leben führen. Ich würde mein Hobby als Beruf ausüben und damit sogar Geld verdienen, es wäre ein Traum. Es gäbe jedoch auch viele Nachteile, ich hätte nicht viel Zeit für meine Frau, was wenn ich Kinder hätte noch schlimmer wäre. Meine Kinder müssten ohne Vater aufwachsen und meine Frau würde sich vielleicht vernachlässigt fühlen.
Ich ~~könnte~~ würde diesen Beruf durchaus ausführen, aber nur wenn ich die richtige Frau hätte die es lange Zeit ohne mich aushalten würde und keine Kinder. Nach meiner Fussballkarriere mit etwa 35 Jahren wäre ich dann auch bereit Kinder zu haben, denn meine Kinder sollten nichts mit den Medien zu tun haben bis sie 18 Jahre alt sind.

* meine Freunde und Verwandten

Textbeispiel 4: 501 bis 600 Punkte

Ein Mittel um sich zu orientieren

Noten werden überall auf unserem Planeten von den Lehrkräften gegeben. Doch sind sie wirklich notwendig und was bringen sie einem?

Ich gehe mittlerweile schon seit acht Jahren in die Schule und bekomme nach jedem absolvierten Semester eine Beurteilung meiner Leistungen. Manchmal ist sie gut und manchmal weniger, doch sie alle haben einen Nutzen. Sie helfen einem, sich zu orientieren. Man weiss so, ganz genau wo man steht und wie man sich in Zukunft verhalten muss, damit man seine Schwächen ausgleichen kann. Sie spielen auch eine grosse Rolle, wenn man später in das Berufsleben einsteigt.
Doch Noten haben auch ihre schlechten Seiten, denn manche Menschen differenzieren sich nur darüber und so kann man sehr schnell ein

Opfer von Mobbing werden, wenn man ein gewisses Ziel nicht erreicht hat. Auch wenn man sich bewirbt sollte der Lehrherr nicht all zu viel Wert auf die Noten legen, denn ein Bewerbungsgespräch sagt deutlich mehr aus.

Meiner Meinung nach sind Noten nicht für die Katz, denn man kann sich so besser orientieren. Schwächen können so besser realisiert und kompensiert oder gar ausgemerzt werden.

Textbeispiel 5: 601 bis 700 Punkte

Unsere heutige Welt funkelt nur von Stars. Überall sieht man berühmte Sänger, Schauspieler, Musiker oder Sportler. Nach aussen wirken sie immer glücklich und lächeln, aber sind sie wirklich so glücklich? Ist das Leben als Berühmtheit erstrebenswert, oder lässt man es lieber sein?

Wäre ich ein Star, würde ich am Liebsten eine Sängerin sein. Morgens, nach einem ausgiebigem Frühstück, ginge ich in mein Tonstudio und überlegte mir wie ein neuer Hit von mir aussehen könnte. Er müsste natürlich so toll sein, dass er sofort auf den ersten Platz der Charts käme. Nach einigen Wochen ginge ich meinen Fans zuliebe auf Welttournee und gäbe in jedem erdenklichem Land ein oder sogar mehrere Konzerte. Aber wenn ich abends einmal kein Konzert geben müsste, wäre mir sicher nicht langweilig, denn ich wäre schon zum nächsten prominenten Anlass eingeladen.

Was mir an einem Leben als Star gefallen würde, wäre der Lebensstil. Die Arbeit ist ein eigenes Hobby und es ist toll, wenn man durch etwas das man liebt, so viel Geld verdienen kann. Man wird zu diversen Events eingeladen und hoch angesehen. Der Grund warum ich jedoch niemals ein Star werden möchte, ist, dass man fast keine Privatsphäre hat. Alles was man macht, weiss gleich die ganze Welt, die ganze Zeit wird man von nervigen Paparazzi oder Journalisten verfolgt und falls sich jemand einen Spass erlaubt, und ein falsches Gerücht verbreitet, muss man sich auch noch den Medien gegenüber rechtfertigen.

Das Leben als Promi mag schön und gut sein, doch für mich wäre das nichts. Es ist super, dass es Stars gibt, denn ohne sie würde viel Unterhaltungs-programm wegfallen, doch ob diese Menschen mit ihrem Leben glücklich sind, wissen nur sie selber.

Textbeispiel 6: 701 bis 800 Punkte

Fast alle haben Angst vor Prüfungen, weil es am Schluss eine Note gibt und diese ins Zeugnis fliesst. Das Zeugnis spielt eine wichtige Rolle für die Zukunft, denn bei Bewerbungen muss man das Zeugnis mitschicken. Auch bei der Gymiprüfung zählen die Vornoten, deshalb will man bei Tests gut abschneiden.
Ich mache gute Erfahrungen mit Noten und bin selten unter einem 4er. Ich freue mich immer darüber, wenn ich eine gute Note habe.
Es gibt viele Vorteile von Noten. Wenn man bei einer Prüfung schlecht abgeschnitten hat, weiss man, dass dieses Thema noch einmal genau angeschaut werden muss und vielleicht Nachilfestunden nützen würden. Gute Noten machen glücklich und man hat ein gutes Zeugnis und wenn die Eltern die guten Ergebnisse sehen, unterschreiben sie mit grosser Freude und sind stolz auf ihr Kind. Ein Vorteil ist auch, dass die Schüler den Stoff gut lernen, um gute Tests zu schreiben. Würde es keine Noten und somit keine Zeugnisse geben, wäre vielen Schülern der Stoff egal und sie würden sich keine Mühe geben. Gute Zeugnisse schinden bei Bewerbungen einen guten Eindruck und bei Gymiprüfungen sind

sie von Nützen. Ich habe die Gymiprüfung auch wegen den Vornoten bestanden. Leider habe ich die Probezeit aber nicht bestanden.
Nachteile von den Noten sind, wenn man eine Zeit lang eine schlechte Phase hat und nur schlechte Prüfungen schreibt. Man kann nach einer schlechten Note in einem Fach Prüfungsangst vor dem nächsten Test haben und schlaflose Nächte erleiden. Nach einer Prüfung mit vielen Fehlern muss man viel Zeit investieren, um diese zu verbessern. Ein Nachteil von Noten ist, dass man sehr viel Zeit mit dem Lernen verbringen muss. An vielen Wochenenden und an Mittwochnachmittagen habe ich viel gelernt und hatte fast keine Freizeit. Wenn Eltern schlechte Noten sehen, können sie einem verbieten, mit Freunden rauszugehen oder das Handy wegnehmen, damit die Kinder mehr Zeit zum Lernen haben. Mir ist das zum Glück noch nie passiert. (Abschnitt)
Die Vorteile zählen für mich mehr als die Nachteile, denn wenn ich gute Noten habe, bin ich glücklich und ich bin lieber draussen, als zu Hause zu Lernen. Ich sehe es immer gerne, wenn ich eine gute Note habe und meine Eltern stolz auf mich sind.

# 9 Fazit

Im Dezember 2012 wurde im Kanton Zürich flächendeckend das Stellwerk-Modul „Texte schreiben" als Ergänzung zu den computerbasierten Stellwerk-Tests durchgeführt. Mit diesem Modul werden die produktiven Fähigkeiten im Fachbereich Deutsch erfasst. Insgesamt nahmen 9953 Schülerinnen und Schüler der 8. Klasse teil.

Die Schülerinnen und Schüler konnten zwischen zwei verschiedenen Themen wählen: 1) „Das Leben als Star" und 2) „Muss es in der Schule Noten geben?". Bei beiden Themen wurden die Schülerinnen und Schüler aufgefordert, zu einem Sachverhalt Stellung zu nehmen sowie die Vor- und Nachteile darzustellen. Drei Fragen dienten den Schülerinnen und Schülern als Leitfaden für das Verfassen der Texte. Ebenfalls zum Schreibauftrag gehörten Angaben darüber, wie die Texte bewertet wurden.

Wie die Analysen zeigten, eignet sich das eingesetzte Beurteilungsraster gut zur Bewertung der Fähigkeiten im Bereich „Texte schreiben". Die Beurteilungskriterien sind bis auf wenige Ausnahmen sehr trennscharf. Das heisst, dass die eingesetzten Beurteilungskriterien wesentlich dazu beitragen, die Fähigkeiten der Schülerinnen und Schüler zu bestimmen und zwar so, dass gute Schülerinnen und Schüler eine hohe Punktzahl erreichen und schwache Schülerinnen und Schüler eine tiefe Punktzahl.

Die Beurteilungsübereinstimmung der korrigierenden Personen (Rater) wurde mit verschiedenen Verfahren überprüft. Die Übereinstimmung zwischen den Ratern im Rahmen der Doppelkorrekturen fiel bei einzelnen Items etwas geringer aus als erwartet. Allerdings liegt diese Abweichung zum grössten Teil im Rahmen von plus/minus einem Punkt pro Kriterium. Zudem hat die bisherige Forschung gezeigt, dass es trotz intensiver Schulung und Betreuung der Rater schwierig ist, eine sehr hohe Übereinstimmung zwischen den Ratern zu erzielen (O'Sullivan & Rignall, 2007)[14].

Es zeigte sich im Rahmen des Moduls „Texte schreiben" dennoch, dass mit einem standardisierten Beurteilungsraster relativ einheitlich beurteilt werden kann. Die Beurteilungen können zudem aus testtheoretischer Perspektive als zuverlässig bezeichnet werden. Mit Hilfe der Item-Response-Theorie konnten die leicht unterschiedlichen Beurteilungsmassstäbe der Rater bei der Berechnung der Ergebnisse der Schülerinnen und Schüler korrigiert werden, so dass eine faire Beurteilung möglich wurde.

Die Ergebnisrückmeldung auf der transformierten Skala (Mittelwert = 500 Punkte, Standardabweichung = 100 Punkte) darf nicht darüber hinwegtäuschen, dass die Ergebnisse unabhängig von den anderen Testergebnissen der Stellwerk-Tests zu interpretieren sind. Der Mittelwert und die Standardabweichung beziehen sich ausschliesslich auf die 9953 beteiligten Schülerinnen und Schüler der 8. Klasse im Kanton Zürich. Die Mädchen erreichen im Durchschnitt signifikant bessere Ergebnisse als die Knaben. Die Darstellung der Ergebnisse nach Schultypen zeigt zudem, wie sinnvoll eine schultyp-unabhängige Beurteilung sein kann. Beispielsweise schreiben rund 10 Prozent der Schülerinnen und Schüler der Sek B Texte, welche in ihrer Beurteilung über dem Mittelwert der Sek A liegen.

Abschliessend kann gesagt werden, dass die Erfassung der produktiven Kompetenzen im Fachbereich Deutsch aus einer testtheoretischen Perspektive als zuverlässig betrachtet werden darf. Wenn standardisierte Beurteilungskriterien sowie eine ausführliche Einarbeitungsphase der Rate-

[14] O'Sullivan, B. & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Modul. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446-478). Cambridge, UK: Cambridge University Press.

rinnen und Rater mit einer klaren Definition des gemeinsamen Beurteilungsmasstabs eingehalten werden, kann die Erfassung von produktiven Kompetenzen mit der Erfassung von reproduktiven Kompetenzen mittels Leistungstests mithalten.

# 10 Fazit der Fokusgruppe

Am 4. Juli 2013 wurde eine Fokusgruppe zum Stellwerk-Modul „Texte schreiben" durchgeführt, bestehend aus dem Volksschulamt, dem Institut für Bildungsevaluation (IBE), Vertreterinnen und Vertretern der Lehrerverbände sowie einzelnen Lehrpersonen. Nachfolgend sind die wichtigsten Aspekte der Diskussion der Fokusgruppe zusammengefasst. Die Anregungen wurden vom Volksschulamt und vom Institut für Bildungsevaluation für die Durchführung des Moduls im Jahr 2013 geprüft und werden wenn möglich miteinbezogen.

### Durchführung

Die Durchführungsdauer von zwei Tagen kann dazu führen, dass gewisse Schülerinnen und Schüler aus benachbarten Sekundarschulen die Themen bereits in Erfahrung bringen und sich somit vor der Durchführung Gedanken über die Themen machen könnten. Dieselbe Problematik stellt sich für Schülerinnen und Schüler, welche an beiden Durchführungstagen krank sind und ihren Text zu einem späteren Zeitpunkt verfassen. Eine mögliche Lösung könnte sein, dass drei Paare von Themen vorbereitet werden, welche zufällig auf die Schulen verteilt werden. Diese Alternative wurde geprüft, allerdings überwiegen die Nachteile. Die technische Durchführung präsentiert sich als äusserst komplex und die Korrekturqualität bei vier bis sechs verschiedenen Themen würde abnehmen. Zusätzlich würde der Aufwand für die Schulung der Korrektor/-innen und der Korrekturaufwand mit sechs verschiedenen Themen deutlich zunehmen. Darüber hinaus zeigt die Erfahrung, dass die Texte der Schülerinnen und Schüler qualitativ nicht besser ausfallen, wenn das Thema im Voraus bekannt ist. Daher wurde entschieden, das bisherige Verfahren beizubehalten und allen Schülerinnen und Schülern dieselben Themen vorzulegen.

### Aufgabenstellung

Grundsätzlich wurden die beiden Themen als ansprechend und lebensnah beurteilt. Das Thema „Das Leben als Star" sei besonders attraktiv für die Schülerinnen und Schüler und lässt einen gewissen Spielraum für Kreativität offen. Es sei jedoch auch gut, dass mit dem Thema „Muss es in der Schule Noten geben?" ein alternatives Thema zur Auswahl gegeben wurde. Die Aufgabenstellung für die Schülerinnen und Schüler zum Verfassen der Texte könnte in einzelnen Punkten noch etwas präziser und transparenter formuliert werden.

### Kriterienraster

Bei der Dimension „Sprachrichtigkeit" wird angeregt, die einzelnen Kriterien differenzierter zu beurteilen, indem anstatt einer dreistufigen eine vierstufige Skala verwendet wird.

### Rückmeldung an die Lehrpersonen

Von einzelnen Lehrpersonen wurde angemerkt, dass Schülerinnen und Schüler im Stellwerk bei der Schreibkompetenz im Vergleich zu den anderen Stellwerk-Tests im Fach Deutsch besser abgeschnitten haben. Die Ergebnisse aus dem Modul „Texte schreiben" wurden mit den computerbasierten Stellwerk-Tests im Fach Deutsch verglichen. Dabei zeigt sich, dass basierend auf der im Kanton St. Gallen durchgeführten Normierung, die Schülerinnen und Schüler im Durchschnitt beim Modul „Texte schreiben" – im Vergleich zu den computerbasierten Tests – tatsächlich höhere Punktzahlen erreichten.

Die Ergebnisrückmeldung an die Lehrpersonen mit den Kompetenzbeschreibungen und Textbeispielen fand bei den Lehrpersonen grossen Anklang. Vermutlich wurden jedoch die Zusatzbeispiele auf der Homepage des IBE nur von wenigen Lehrpersonen beachtet. Wären sie allerdings nicht zur Verfügung gestellt worden, hätte dies zu Unmut führen können.

### Bewertung der Aufsätze

Es wurde angemerkt, dass einige Schülerinnen und Schüler beim Modul „Texte schreiben" besser abschnitten als die Lehrpersonen erwartet hätten. Es wird vermutet, dass dies darauf beruhen könnte, dass bei Sek-A-Schüler/-innen die Sprachrichtigkeit und die Sprachangemessenheit im Schulalltag im Vergleich zum Modul „Texte schreiben" stärker gewichtet werden. Hierbei muss jedoch beachtet werden, dass die Lehrerbeurteilung auf den Leistungen einer Klasse basiert, während den Ergebnissen des Moduls „Texte schreiben" die Leistungen aller Schülerinnen und Schüler der 8. Klasse im Kanton Zürich zugrunde liegen. Dies führt zu unterschiedlichen Bewertungsmassstäben. Das Ergebnis des Moduls „Texte schreiben" ist aufgrund seiner standardisierten Konzipierung als valider zu bewerten.

Grundsätzliche Bemerkung zum Modul: Der standardisierte Test „Texte schreiben" wird insgesamt als gut bewertet. Es stellt sich jedoch die Frage, welche Instanz (das IBE, Fachdidaktiker, die Bildungspolitik) definiert, was gute Schreibkompetenzen sind.