Evaluation zum neuen Übertrittsverfahren in die Maturitätsschulen im Kanton Zürich basierend auf bildungsstatistische Daten und Befragungen von Schülerinnen und Schülern: Schlussbericht zuhanden der Bildungsdirektion

Martin J. Tomasik

Universität Zürich

Institut für Erziehungswissenschaft

Methoden der Entwicklungs- und Bildungsforschung

Impressum

Publikationsdatum
2. Juni 2026

Herausgeberin
Kanton Zürich
Bildungsdirektion Bildungsplanung
Walcheplatz 2
8090 Zürich

https://www.zh.ch/de/bildungsdirektion/generalsekretariat-der-bildungsdirektion/bildungsplanung.html

Autor
Prof. Dr. Martin J. Tomasik
Universität Zürich
Institut für Erziehungswissenschaft
Methoden der Entwicklungs- und Bildungsforschung
Kantonsschulstrasse 3
8001 Zürich

www.ife.uzh.ch

Zitationsvorschlag
Tomasik, M. J. (2026). *Evaluation zum neuen Übertrittsverfahren in die Maturitätsschulen im Kanton Zürich basierend auf bildungsstatistische Daten und Befragungen von Schülerinnen und Schülern*. Universität Zürich.

Gesamtinhaltsverzeichnis

Vorwort

Bildungsdirektion

Die Zürcher Maturitätsschulen sind beliebt, viele Schülerinnen und Schüler der Volksschule melden sich für die Prüfung zu einer weiterführenden Schule an. Über den Zugang zu diesen Schulen entscheidet das Übertrittsverfahren im Kanton Zürich. Es ist darum wichtig, dass das Verfahren für alle Beteiligten – Schülerinnen und Schüler, Eltern, Lehrpersonen und Schulbehörden – verlässlich und verständlich ist.

Mit der Verordnung über das Aufnahmeverfahren an Mittelschulen (VAM) sowie den ergänzenden Anpassungen im Aufnahmereglement für das Langzeitgymnasium (AufnahmeR LG) hat der Kanton Zürich das Übertrittsverfahren in den letzten Jahren weiterentwickelt. Die verschiedenen Übertrittsverfahren wurden harmonisiert, die Prüfungsinhalte und Bestehensnormen vereinheitlicht, schulische Vorleistungen systematisch einbezogen und der administrative Aufwand reduziert. Im Kanton Zürich gilt das vereinheitlichte Verfahren seit dem Schuljahr 2022/23.

Welche Auswirkungen bringt das neue Übertrittsverfahren nun mit sich? Um dieser Frage nachzugehen, führte die Professur für Methoden der Entwicklungs- und Bildungsforschung am Institut für Erziehungswissenschaft der Universität Zürich im Auftrag der Bildungsdirektion eine Evaluation durch. Im Fokus standen folgende Fragen: Welche Veränderungen gehen mit den Anpassungen im Übertrittsverfahren einher? Und wie wird die Zentrale Aufnahmeprüfung (ZAP) in ihrer neuen Ausrichtung von den Schülerinnen und Schülern wahrgenommen?

Der vorliegende Schlussbericht stützt sich auf bildungsstatistische Auswertungen und auf eine Onlinebefragung von über 3 000 Jugendlichen, die eine Aufnahmeprüfung abgelegt haben. Die Verbindung dieser Daten zeigt ein differenziertes Bild zu den Auswirkungen des neuen Übertrittsverfahrens. Die Auftragnehmenden fokussieren in ihrem Bericht auf die Beschreibung der Ergebnisse. Die Bewertung der Ergebnisse übernimmt die Bildungsdirektion in ihrem ergänzenden Evaluationsbericht, in welchem sie die Befunde einordnet und Handlungsfelder identifiziert.

Ein grosser Dank gebührt dem Forschungsteam der Universität Zürich für seine methodisch fundierte Arbeit, allen teilnehmenden Jugendlichen sowie allen weiteren Personen, welche die Evaluation unterstützt haben.

Dr. Sybille Bayard
Leiterin Bildungsplanung
Bildungsdirektion des Kantons Zürich

Kapitel 1
Bildungsstatistische Auswertungen

Martin J. Tomasik

Inhaltsverzeichnis

1. Einleitung

Die bildungsstatistischen Auswertungen erfolgen anhand von Leitfragen, die von der Auftraggeberin definiert worden sind und die zentralen Erkenntnisinteressen dieses Teils der Evaluation zusammenfassen. Die folgenden Leitfragen strukturieren die bildungsstatistischen Analysen in Kapitel 1; die jeweiligen Antworten finden sich in den angegebenen Abschnitten.

- Wie unterscheiden sich die Prüfungsnoten nach Gruppenzugehörigkeit (Geschlecht, Nationalität, Muttersprache, Dauer in der Schweiz, Trägerschaft der abgebenden Schule? (vgl. Abschnitt 3 und Abschnitt 3.1)
- Wie stark ist der statistische Zusammenhang zwischen Vornoten und Noten der Aufnahmeprüfung? (vgl. Abschnitt 4)
- Mit welchen Vornoten gehen die Jugendlichen an die Prüfung? Wie stark streut dies zwischen Schulgemeinden, Schulen, Klassen und Jugendlichen? (vgl. Abschnitt 5)
- Welche Vornoten haben bestandene und abgelehnte Jugendliche? (vgl. Abschnitt 6)
- Wie gross ist der Anteil der Kandidat*innen, die die Prüfung bestanden hätten, wenn die Vornote nicht gezählt hätte? (vgl. Abschnitt 7)
- Wie haben sich die Anmeldezahlen, Bestehenszahlen und Bestehensquoten mit VAM/AufnahmeR LG entwickelt? (vgl. Abschnitt 8)
- Wie haben sich die Anmeldezahlen und Bestehensquoten nach Gruppenzugehörigkeit mit VAM/AufnahmeR LG verändert? (vgl. Abschnitt 9)
- Welcher Anteil der Jugendlichen hat sich bei der Anmeldung für die KG-Prüfung auch für die HMS-Option entschieden? (vgl. Abschnitt 10)
- Wie haben sich Anmelde- und Bestehensquoten bei Doppelanmeldungen KG/HMS mit VAM/AufnahmeR LG entwickelt? Hat sich dieser Zusammenhang mit VAM/AufnahmeR LG geändert? (vgl. Abschnitt 10)

Für die Beantwortung dieser Leitfragen ist in den meisten Fällen ein Vergleich zwischen dem Prüfungsregime ante-VAM (2020-2022) mit dem Prüfungsregime post-VAM (2023-2024) notwendig,[1] wobei als Datengrundlage für das erste die Prüfungen der Jahre 2020, 2021 und 2022 und als Datengrundlage für das zweite die Prüfungen der Jahre 2023 und 2024 berücksichtig werden. Dadurch ist die Datengrundlage post-VAM (2023-2024) etwas schmaler als die Datengrundlage ante-VAM (2020-2022) und jahrgangsspezifische Effekte lassen sich somit nicht gänzlich ausschliessen.

Sämtliche Auswertungen wurden getrennt nach den Prüfungstypen LG (Langgymnasium), KG (Kurzgymnasium), HMS (Handelsmittelschule), FMS

[1] Zur besseren Lesbarkeit wird im Folgenden von «ante-VAM» und «post-VAM» gesprochen, wobei damit auch der Regimewechsel vor und nach der Reform des AufnahmeR LG mitgemeint ist.

(Fachmittelschule), IMS (Informatikmittelschule) und BM1 (Berufsmaturitätsschule beim Besuch während der beruflichen Grundbildung) durchgeführt. In diesem Abschlussbericht liegt der Fokus wegen der bildungspolitischen Bedeutsamkeit dieser Prüfungstypen auf dem LG- und dem KG-Prüfungstyp. In der Regel wird auf Ergebnisse zu anderen Prüfungstypen nur dann eingegangen, wenn wesentliche Unterschiede zu diesen beiden auffällig sind. [...]

2. Methodische Vorbemerkungen

Grundsätzlich ist es so, dass in allen folgenden Auswertungen die Untersuchungseinheit immer eine einzelne Prüfung und nicht eine einzelne Person ist. Eine Person kann also mehrfach auftauchen, wenn sie sich für mehrere Prüfungen in einem Jahr anmeldet (und diese dann mit eindeutigem Ergebnis besteht oder nicht besteht) oder wenn sie sich in verschiedenen Jahren für die gleiche oder eine andere Prüfung anmeldet (und diese dann mit eindeutigem Ergebnis besteht oder nicht besteht). Wenn also im Folgenden von Jugendlichen oder Kandidat*innen die Rede ist, dann ist das immer so gemeint, dass eine Person in einer bestimmten Prüfung berücksichtigt wird.

Mit anderen Worten kann man sagen, dass es nicht immer unterschiedliche Personen sind, die gezählt werden. Jemand, der sich ante-VAM (2020-2022) für das LG prüfen liess, diese Prüfung nicht bestand, sich dann für das KG prüfen liess und diese Prüfung dann bestand, taucht also in den Auswertungen doppelt auf – und zwar in diesem Beispiel einmal vor und einmal nach dem Regimewechsel. Wenn diese beispielhafte Person ein Mädchen ist, dann trägt sie (geringfügig) dazu bei, dass ante-VAM (2020-2022) die Mädchen in der LG-Prüfung schlechter abschneiden und post-VAM (2023-2024) in der KG-Prüfung besser.

Für die bildungspolitische Interpretation der Daten spielt das wohl keine grosse Rolle und die konsekutiven «Prüfungskarrieren» sind nicht im Fokus dieser Evaluation. Trotzdem sollte dieser Umstand berücksichtigt werden, wenn man beispielsweise die Gesamtzahl der Anmeldungen (wie sie in dieser Evaluation dargestellt wird) mit der Gesamtzahl der Schüler*innen (wie sie in der Bildungsstatistik gezählt wird) in einen Zusammenhang bringt. Weil mehrfache und wiederholte Prüfungen möglich sind, ist der Anteil der Schüler*innen, die sich prüfen lassen, gemessen an allen Schüler*innen im Kanton nicht ganz so gross, wie das auf den ersten Blick aussieht. Für weitere Einschränkungen bei der Interpretation der Gesamtzahlen siehe auch die Ausführungen zur Definition der Population weiter unten.

Die Population wurde so definiert, dass entweder ein eindeutig positiver oder ein eindeutiger negativer Zulassungsentscheid vorgelegen hat (wobei der Prüfungstyp BM2 von vornherein ausgeschlossen wurde). Alle anderen Fälle wurden aus der Datenbank entfernt. Das beinhaltet beispielsweise Jugendliche mit fehlenden Prüfungsnoten, solche die aus welchen Gründen auch immer prüfungsfrei zugelassen worden sind, solche die eine mündliche Nachprüfung absolviert haben usw.

Das hat Konsequenzen für die Interpretation der Daten. So war es beispielsweise ante-VAM (2020-2022) möglich, bei einer knapp nicht bestandenen schriftlichen Prüfung eine mündliche Nachprüfung zu absolvieren. Weil die knapp nicht bestandene schriftliche Prüfung kein eindeutig positives oder eindeutig negatives Prüfungsergebnis darstellt, wird es in den folgenden Auswertungen überhaupt nicht berücksichtigt. Es zählt vielmehr das Ergebnis der mündlichen Nachprüfung. Dieses Vorgehen ist einerseits notwendig, um die Komplexität in den Daten zu reduzieren, und andererseits, um solche Kandidat*innen nicht doppelt zu zählen. Ausserdem ist so die Vergleichbarkeit zwischen ante-VAM (2020-2022) und post-VAM (2023-2024) besser gewährleistet.

Das Alter zum Zeitpunkt der Prüfung (und damit auch die Dauer des Aufenthalts in der Schweiz, falls die Angabe «seit Geburt» erfolgte) wurde unter der Annahme berechnet, dass die Prüfung am 15. März eines jeweiligen Jahres stattgefunden hat.

Die Nationalitäten wurden in den Kategorien «Schweiz» (CHE), «Europa ex Schweiz» (einschliesslich Russland, ohne Türkei) und «Sonstige» eingeteilt. In der Kategorie «CHE» wird nicht weiter differenziert, ob neben der Schweizer auch noch eine weitere Nationalität gegeben ist oder nicht. Zwar zeigen sich hier zuweilen interessante Unterschiede zwischen diesen Gruppen, aber die Unterscheidung wurde erst seit 2022 erfasst und kann somit nicht für die Vergangenheit übernommen werden.

Für die Muttersprache wurden vier Kategorien unterschieden. «Deutsch» (DEU) wurde als grösste Kategorie einzeln geführt. Sonstige germanische Sprachen (z. B. Englisch oder Norwegisch) sowie romanische Sprache (z. B. Portugiesisch oder Rumänisch) wurden der Kategorie «sonstige germanische und romanische Sprachen» (SGR) zugeschlagen. Alle anderen europäischen Sprachen (z. B. Polnisch oder Griechisch) wurden unter «nicht germanisch oder romanisch» (NGR) zusammengefasst. Alle weiteren Sprachen (z. B. Chinesisch oder Arabisch) gingen in die Restkategorie «Sonstige» (ETC) ein.

Bei der Trägerschaft wurde nach öffentlichen und privaten Trägern unterschieden. Diese Angaben stammen direkt aus der Bildungsstatistik, die von der Bildungsdirektion zur Verfügung gestellt worden ist. Neben der Trägerschaft der abgebenden Schule wurde auch der abgebende Schultyp aus der Bildungsstatistik identifiziert und in die Kategorien «Primarschule», «Sekundarschule A» und «Sonstige» kodiert. Da diese Variable sehr hoch mit dem Prüfungstyp konfundiert ist, wurde sie in den weiteren Auswertungen nicht weiter berücksichtigt und taucht nur in diesem Abschnitt bei den deskriptiven Auswertungen einmal auf.

Die Aufenthaltsdauer (eigentlich: Nutzung der deutschen Sprache) wurde in drei Gruppen unterteilt, nämlich «einheimisch», «bis zu 10 Jahre» und «mehr als 10 Jahre», wobei die letzte Gruppe auch solche Fälle enthält, die angegebene haben, seit Geburt Deutsch zu sprechen. Die Zehn-Jahres-Grenze wurde deswegen gezogen, da erstens ab zehn Jahren keine Differenzierung mehr in den Daten stattfindet und dies zweitens für viele Ausländer*innen der Zeitpunkt ist, zu dem sie sich einbürgern lassen könnten. In die Kategorie «einheimisch» wurde auch aufgenommen, wer auf dieser Variable keine Angabe

gemacht hat. Aus diesem Grund wird die Grösse dieser Gruppe vermutlich leicht überschätzt.

Es wurden 7’390 Kandidat*innen ausgeschlossen, bei denen nach der obigen Definition kein eindeutiges Prüfungsergebnis vorlag. Die meisten Ausschlüsse gab es für den sowieso nicht berücksichtigten Prüfungstyp BM2 (3’934 Ausschlüsse) sowie für das KG (1’630 Ausschlüsse). Des Weiteren wurden 287 Kandidat*innen ausgeschlossen, für die keine gültige Angaben zum aufnehmenden Schultyp vorlagen.

Es verblieben *N* = 50’992 Jugendliche (bzw. genauer gesagt einzelne eindeutig als bestanden oder nicht bestanden auswertbare Prüfungen) als Grundgesamtheit für alle folgenden Auswertungen, die sich wie in **Tabelle** dargestellt auf die einzelnen Prüfungstypen verteilen.

Tabelle 1

Gesamtheit der analysierten Prüfungen vor und nach dem Regimewechsel, aufgeteilt nach Prüfungstyp

Verteilung der Prüfungstypen

Prüfungstyp	N
LG	22511
KG	17979
HMS	1147
FMS	3170
IMS	600
BM1	5585
Total	50992

Die Verteilungen der einzelnen Gruppenvariablen finden sich in **Tabelle 2**. Gelegentlich fehlen Ausprägungen auf einzelnen Variablen, weshalb das Total sich von Tabelle zu Tabelle unterscheidet. Die Angabe zur Trägerschaft fehlt bei etwa der Hälfte der Stichprobe. Bei manchen Variablen (wie etwa der Aufenthaltsdauer in der Schweiz) existiert die Variable nur für eine Untergruppe, die dann deutlich kleiner sein kann. Für das Geschlecht, die Nationalität, die Muttersprache und den abgebenden Schultyp liegen vollständige Daten vor.

Über die Zeit hinweg könnte es Änderungen bei der Operationalisierung der verschiedenen Variablen gegeben haben, zum Beispiel dadurch, dass sich die Formulierung von Fragen im Fragebogen zur Prüfungsanmeldung geändert hat. Dies würde den Vergleich zwischen den einzelnen Jahren und den Vergleich zwischen ante-VAM (2020-2022) und post-VAM (2023-2024) erschweren. An dieser Stelle wird überprüft, inwieweit die Verteilungen der Gruppenvariablen über die Zeit hinweg einigermassen

stabil sind. Geringfügige Änderungen kann es geben, da es ja auch tatsächlich Veränderungen in der Bevölkerungszusammensetzung gibt.

Insgesamt zeigte sich eine hohe Stabilität der Gruppenvariablen über die Jahre. Die Zusammensetzung der Geschlechter unterscheidet sich um maximal 2 Prozentpunkte und die der Nationalität um maximal 1 Prozentpunkt. Dafür nimmt der Anteil der Kandidaten mit Muttersprache Deutsch über die drei untersuchten Jahre um etwa 5 Prozentpunkte ab. Im Gegenzug steigt der Anteil der anderen Muttersprachen an, insbesondere bei den sonstigen germanischen und den romanischen Sprachen. Die Aufenthaltsdauer sowie die Verteilung der Trägerschaft variiert in den beiden Jahren, in denen diese Daten überhaupt vorhanden sind, um 1 bis 2 Prozentpunkte.

Tabelle 2

Gesamtheit der analysierten Prüfungen vor und nach dem Regimewechsel, aufgeteilt nach Geschlecht, Nationalität, Muttersprache, Aufenthaltsdauer, Trägerschaft (nur 2022 bis 2024) und abgebenden Schultypen

Verteilung des Geschlechts

Geschlecht	N
weiblich	26467
männlich	24525
	50992

Verteilung der Nationalität

Nationalität	N
Schweiz	41866
Europa ex Schweiz	7050
Sonstige	2076
Total	50992

Verteilung der Muttersprache

Muttersprache	N
DEU	39050
SGR	4978
ETC	3551
NGR	3413
Total	50992

Verteilung der Aufenthaltsdauer (nur 2022/2023)

Dauer	N
bis zu 10 Jahre	3774
mehr als 10 Jahre	4354
einheimisch	42864
Total	50992

Verteilung der Trägerschaft (nur 2022/2023)

Trägerschaft	N
öffentlich	30361
privat	2877
Total	33238

Verteilung der abgebenden Schultypen

Abgebender Schultyp	N
Sekundarschule A	25033
Primarschule	22404
Sonstige	3555
Total	50992

Insgesamt zeigen die Verteilungen der verfügbaren Gruppenvariablen über die Jahre eine hohe Stabilität. Die Zusammensetzung nach Geschlecht variiert maximal um 2 Prozentpunkte und nach Nationalität maximal um 1 Prozentpunkt; auch dort, wo Daten nur für Teil-Zeiträume vorliegen, bewegen sich die Veränderungen (z. B. bei Trägerschaft und Aufenthaltsdauer in den Jahren mit vorhandenen Angaben) im Bereich von 1 bis 2 Prozentpunkten. Vor diesem Hintergrund lassen sich gruppenspezifische Veränderungen in den folgenden Auswertungen in der Regel direkt zwischen den Jahren bzw. im Vergleich ante-VAM (2020-2022) vs. post-VAM (2023-2024) betrachten, ohne dass eine aufwändige Korrektur entlang der Bevölkerungszusammensetzung erforderlich ist. Gleichzeitig ist bei

der Interpretation konsequent zu berücksichtigen, dass einzelne Variablen (insbesondere die Trägerschaft) häufig fehlen und dass bestimmte Merkmale (wie etwa die Aufenthaltsdauer) nur für Teilgruppen erfasst sind; dadurch können sich die Bezugsgrössen zwischen Tabellen und Abbildungen unterscheiden. Vergleiche sollten daher primär innerhalb derselben Datengrundlage (gleiche Variable, gleiche Teilmenge, gleiche Jahrgänge) gezogen werden, und Unterschiede über verschiedene Teilmengen hinweg sind zurückhaltend zu interpretieren (Schafer & Graham, 2002). Als inhaltlich relevante Ausnahme ist die Muttersprache hervorzuheben: Der Anteil der Kandidat*innen mit Muttersprache Deutsch ist beispielsweise im Zeitraum 2020 bis 2023 um rund 5 Prozentpunkte gesunken (von 80 auf 75 Prozent). Dieser Kontext ist bei Befunden zu sprachbezogenen Gruppenunterschieden mitzudenken, da solche Verschiebungen in der Zusammensetzung die Einordnung von gruppenspezifischen Kennwerten beeinflussen können, ohne dass damit eine Veränderung der Prüfungsprozesse oder der Leistungsniveaus impliziert ist.

3. Prüfungsdurchschnitte vor und nach dem Regimewechsel

Zunächst wurde überprüft, wie sich die Prüfungsnoten der unterschiedlichen Übertrittsprüfungen vor und nach dem Regimewechsel verändert haben. Dazu werden im Folgenden standardisierte Differenzen (Cohens *d*) berichtet und zur Einordnung werden 95-Prozent-Konfidenzintervalle angegeben. Verglichen wurden die Prüfungsnoten in Mathematik, in Deutsch (Aufsatz), in Deutsch (Grammatik) sowie die Gesamtnote in Deutsch, die sich aus den beiden Teilnoten in Deutsch ergibt. Mögliche Unterschiede lassen sich dabei auf verschiedene Ursachen zurückführen und dadurch, dass sich in jedem Prüfungsjahr die Skalierung der Prüfung unterschied, lassen sie sich nicht direkt miteinander vergleichen. Trotz der fehlenden Vergleichbarkeit werden hier die Effektstärken [2] berichtet, weil sie eine Grundlage für die Auswertungen nach Gruppenzugehörigkeit bilden, die im folgenden Abschnitt vorgestellt werden. Die Auswertungen nach Gruppenzugehörigkeit sind nämlich wieder möglich, weil die Skalierung für die zu unterscheidenden Gruppen (etwa Jungen und Mädchen) ja dann wieder die gleiche ist.

Insgesamt lässt sich feststellen, dass die Prüfungsnoten für fast alle Prüfungsfächer und Prüfungstypen post-VAM (2023-2024) deutlich höher waren als ante-VAM (2020-2022). Wie in **Tabelle** ersichtlich wird, werden dabei häufig mittlere Effektstärken erreicht. So ist beispielsweise in **Abbildung 1** zu sehen, dass für die LG-Prüfung der

[2] Für die standardisierten Differenzen wird zwecks sprachlicher Vereinfachung im Folgenden immer der technische Begriff der «Effektstärke» benutzt, obgleich es sich bei den zugrunde liegenden Phänomenen nicht um Effekte im Sinne eines Ursache-Wirkungs-Verhältnisses handeln muss, sondern auch um korrelative Zusammenhänge handeln kann.

Prüfungsdurchschnitt in Mathematik von *M* = 3.47 (*SD* = 1.08) ante-VAM auf *M* = 4.03 (*SD* = 1.01) post-VAM gestiegen ist, was einer mittleren Effektstärker von *d* = .53 entspricht. Ähnlich verhält es sich für die KG-Prüfung in Mathematik mit *M* = 3.25 (*SD* = 1.20) ante-VAM und *M* = 4.02 (*SD* = 1.39) post-VAM (*d* = .74) oder für die FMS-Prüfung in Deutsch (Gesamt) mit *M* = 4.05 (*SD* = 0.73) ante-VAM und *M* = 4.58 (*SD* = 0.70) post-VAM (*d* = .74). Dagegen ist beispielsweise der Prüfungsdurchschnitt für die BM1-Prüfung in Mathematik von *M* = 4.24 (*SD* = 1.29) ante-VAM auf *M* = 4.03 (*SD* = 1.33) post-VAM gesunken (*d* = -.16).

Tabelle 3

Standardisierte Mittelwertsdifferenzen bei den Prüfungsnoten in Mathe, Deutsch (Gesamt), Deutsch (Aufsatz) und Deutsch (Grammatik) als Ergebnis des Regimewechsels

Effektstärken des Regimewechsels auf die Prüfungsnote Mathematik

Prüfung	Effektstärke d	CI-	CI+
LG	0.53	0.50	0.56
KG	0.60	0.57	0.63
HMS	-0.17	-0.30	-0.05
FMS	0.56	0.49	0.63
IMS	0.37	0.21	0.53
BM1	-0.16	-0.21	-0.10

Effektstärken des Regimewechsels auf die Prüfungsnote Deutsch (Gesamt)

Prüfung	Effektstärke d	CI-	CI+
LG	0.55	0.52	0.58
KG	0.28	0.25	0.31
HMS	0.20	0.08	0.33
FMS	0.73	0.66	0.81
IMS	0.14	-0.02	0.30
BM1	0.36	0.30	0.41

Effektstärken des Regimewechsels auf die Prüfungsnote Deutsch (Aufsatz)

Prüfung	Effektstärke d	CI-	CI+
LG	0.51	0.49	0.54
KG	0.48	0.45	0.51
HMS	0.51	0.39	0.64
FMS	0.11	0.04	0.18
IMS	0.07	-0.09	0.23

Effektstärken des Regimewechsels auf die Prüfungsnote Deutsch (Grammatik)

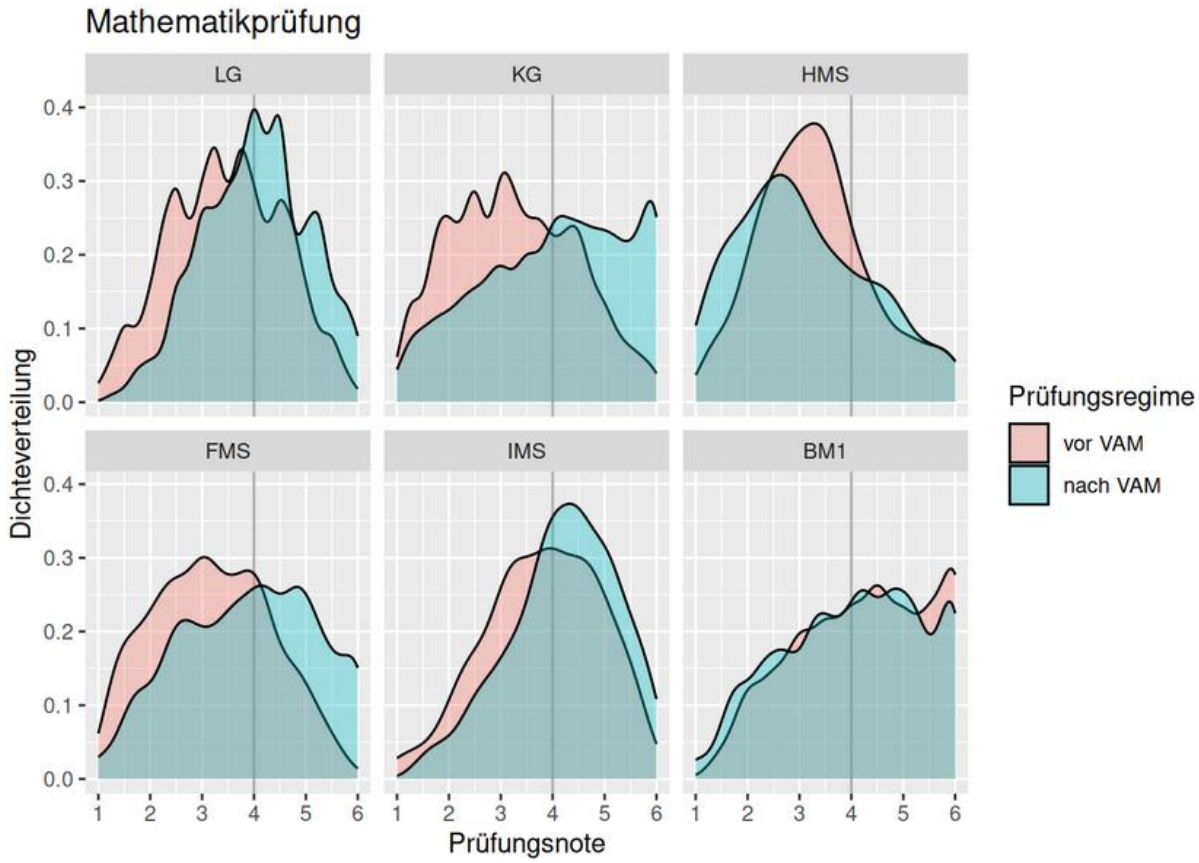
Prüfung	Effektstärke d	CI-	CI+
LG	0.47	0.45	0.50
KG	0.04	0.01	0.07
HMS	-0.12	-0.25	0.00
FMS	1.20	1.12	1.28
IMS	0.18	0.01	0.34

Zur Einordnung der beschriebenen Veränderungen in dieser Tabelle ist zunächst wichtig zu verstehen, dass die Prüfungsnoten zwar über die Jahre auf derselben Notenskala ausgewiesen werden, die zugrunde liegenden Prüfungsjahrgänge jedoch nicht identische Testformen darstellen und sich die Skalierung zwischen den Prüfungsjahren unterscheiden kann. Die hier berichteten Effektstärken (also standardisierte Mittelwertdifferenzen *d*) sind daher primär als deskriptive Zusammenfassung der beobachteten Unterschiede zwischen ante-VAM (2020–2022) und post-VAM (2023-2024) zu verstehen und erlauben ohne zusätzliche Verknüpfung bzw. Equating keine streng

«formgleichen» Aussagen im Sinne eines direkten Vergleichs über unterschiedliche Prüfungseditionen hinweg (Dorans et al., 2010).

Abbildung 1

Verteilung der Prüfungsnoten in Mathematik vor und nach dem Regimewechsel, getrennt dargestellt für die unterschiedlichen Prüfungstypen



Gleichzeitig sind Effektstärken in diesem Kontext eine hilfreiche Verständnishilfe, da sie – im Gegensatz zu reinen Signifikanztests – die Grössenordnung der Verschiebungen auf einer einheitlichen Skala ausdrücken und damit auch bei grossen Stichproben eine inhaltliche Einordnung ermöglichen (Maher et al., 2013). Die in den Abbildungen dargestellten Verteilungen illustrieren diese Verschiebungen anschaulich: Besonders deutlich ist dies beispielsweise bei der KG-Prüfung in Mathematik, wo nach dem Regimewechsel die Häufigkeit von Noten unterhalb einer 4 abnimmt und die Häufigkeit von Noten im Bereich 5 bis 6 zunimmt, während die Effektstärken für andere Prüfungstypen und insbesondere für Deutsch weniger stark ausgeprägt sind. Insgesamt bildet dieser Abschnitt damit den empirischen Ausgangspunkt für die anschliessenden Auswertungen, in denen geprüft wird, ob und in welchem Ausmass sich diese Veränderungen in den Prüfungsdurchschnitten in Abhängigkeit von der jeweiligen Gruppenzugehörigkeit unterscheiden.

3.1 Veränderung der Prüfungsdurchschnitte nach Gruppenzugehörigkeit

Die im letzten Abschnitt beschriebene Veränderung der Prüfungsdurchschnitte infolge des Regimewechsels dient als Grundlage für die Interpretation der folgenden Auswertungen, bei denen diese Veränderungen als Funktion der Gruppenzugehörigkeit modelliert werden. Als Gruppenvariablen wurden dabei das Geschlecht, die Nationalität, die Muttersprache, die Aufenthaltsdauer in der Schweiz sowie die Trägerschaft der abgebenden Schule verwendet. Hinweise zur Operationalisierung dieser Gruppenvariablen finden sich weiter oben.

Insgesamt zeigten sich punktuelle Unterschiede in einzelnen Prüfungen und Fächern, jedoch keine systematischen Gruppenunterschiede. Das spricht dafür, dass durch den Regimewechsel keine Gruppe systematisch bevorzugt oder benachteiligt worden ist. Beispielsweise liegen die *d*-Schätzungen in Mathematik für LG bei Jungen bei $d = 0.59$ (CI 0.54–0.64) und bei Mädchen bei $d = 0.48$ (CI 0.44–0.53); die Intervalle überlappen deutlich, was in den folgenden Tabellen erkennbar ist. Auf Ausnahmen wird an entsprechender Stelle eingegangen, wenn im Folgenden die Unterschiede zwischen Gruppen einzeln diskutiert werden.

Die Effektstärken des Regimewechsels nach Geschlecht finden sich in **Tabelle** getrennt nach den einzelnen Prüfungstypen. Mit zwei Ausnahmen bei der Prüfungsnote Deutsch *(Aufsatz) finden sich keine Unterschiede in der Notenveränderung vor und nach dem* Regimewechsel zwischen Knaben und Mädchen, was dafürspricht, dass der Regimewechsel keines der beiden Geschlechter bevorzugt oder benachteiligt hat. Die erste Ausnahme ist, dass sich bei der Prüfungsnote Deutsch (Aufsatz) in der KG-Prüfung die Knaben mit einer Effektstärke von $d = .56$ (CI 0.50–0.62) vom Regimewechsel stärker verbessert haben als Mädchen mit einer Effektstärke von $d = .44$ (CI 0.38–0.49), wobei die mittlere Effektstärke, wie im letzten Abschnitt gezeigt, bei $d = .49$ liegt. Die zweite Ausnahme ist, dass sich bei der Prüfungsnote Deutsch (Aufsatz) in der IMS-Prüfung die Mädchen mit einer Effektstärke von $d = 1.24$ (CI 0.30–2.15) vom Regimewechsel stärker verbessert haben als die Knaben mit einer Effektstärke von $d = -.14$ (CI -0.35–0.06), wobei die mittlere Effektstärke, wie im letzten Abschnitt gezeigt, bei $d = -.10$ liegt. Dieser Unterschied ist allerdings auf sehr wenige leistungsstarke Mädchen zurückzuführen, wie eine genauere Auswertung der Verteilungen gezeigt hat und was sich an dem sehr breiten Konfidenzintervall (d.h. einer hohen Unsicherheit der Punktschätzung) ablesen lässt. Auch in **Tabelle** gab es auch bezüglich der Nationalität praktisch keine differenziellen Effekte des Regimewechsels). Einzige Ausnahme ist hier die Prüfungsnote in Mathematik bei der KG-Prüfung. Hier haben Jugendliche mit Schweizer Nationalität ($d = .78$; CI 0.72–0.84) nach dem Regimewechsel ihre Noten etwas mehr verbessert als Jugendliche mit sonstiger Nationalität ($d = .45$; CI 0.39–0.51), wobei der Effekt für alle im Durchschnitt bei $d = .74$ liegt. Hinsichtlich der Muttersprache gab es dafür keine auffälligen Unterschiede, wie in Tabelle zu sehen ist.

Tabelle 4

Effektstärken der Notenveränderung nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und nach dem Geschlecht der Kandidat*innen

Effektstärken des Regimewechsels nach Geschlecht auf die Prüfungsnote Mathematik

Prüfung	Geschlecht	Effektstärke d	CI-	CI+
LG	männlich	0.56	0.52	0.60
LG	weiblich	0.50	0.46	0.54
KG	männlich	0.60	0.55	0.64
KG	weiblich	0.61	0.57	0.65
HMS	männlich	-0.16	-0.32	0.01
HMS	weiblich	-0.25	-0.45	-0.06
FMS	männlich	0.59	0.45	0.73
FMS	weiblich	0.54	0.45	0.63
IMS	männlich	0.38	0.21	0.55
IMS	weiblich	0.11	-0.52	0.75
BM1	männlich	-0.17	-0.25	-0.10
BM1	weiblich	-0.14	-0.23	-0.06

Effektstärken des Regimewechsels nach Geschlecht auf die Prüfungsnote Deutsch (Gesamt)

Prüfung	Geschlecht	Effektstärke d	CI-	CI+
LG	männlich	0.57	0.53	0.60
LG	weiblich	0.55	0.52	0.59
KG	männlich	0.26	0.21	0.30
KG	weiblich	0.30	0.26	0.34
HMS	männlich	0.22	0.05	0.38
HMS	weiblich	0.22	0.02	0.41
FMS	männlich	0.83	0.69	0.97
FMS	weiblich	0.71	0.63	0.80
IMS	männlich	0.13	-0.04	0.30
IMS	weiblich	0.33	-0.32	0.97
BM1	männlich	0.39	0.32	0.47
BM1	weiblich	0.33	0.24	0.41

Effektstärken des Regimewechsels nach Geschlecht auf die Prüfungsnote Deutsch (Aufsatz)

Prüfung	Geschlecht	Effektstärke d	CI-	CI+
LG	männlich	0.55	0.51	0.59
LG	weiblich	0.50	0.46	0.54
KG	männlich	0.51	0.47	0.56
KG	weiblich	0.47	0.43	0.51
HMS	männlich	0.52	0.35	0.68
HMS	weiblich	0.54	0.35	0.74
FMS	männlich	0.17	0.04	0.31
FMS	weiblich	0.10	0.02	0.19
IMS	männlich	0.05	-0.12	0.22
IMS	weiblich	0.41	-0.24	1.05

Effektstärken des Regimewechsels nach Geschlecht auf die Prüfungsnote Deutsch (Grammatik)

Prüfung	Geschlecht	Effektstärke d	CI-	CI+
LG	männlich	0.46	0.43	0.50
LG	weiblich	0.49	0.45	0.53
KG	männlich	-0.02	-0.07	0.02
KG	weiblich	0.09	0.05	0.13
HMS	männlich	-0.11	-0.27	0.05
HMS	weiblich	-0.11	-0.31	0.08
FMS	männlich	1.31	1.16	1.46
FMS	weiblich	1.17	1.08	1.26
IMS	männlich	0.18	0.01	0.35
IMS	weiblich	0.18	-0.46	0.82

Tabelle 5

Effektstärken der Notenveränderung nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und nach der Nationalität der Kandidat*innen

Effektstärken des Regimewechsels nach Nationalität auf die Prüfungsnote Mathematik

Prüfung	Nationalität	Effektstärke d	CI-	CI+
LG	Schweiz	0.53	0.50	0.56
LG	Europa ex Schweiz	0.56	0.49	0.63
LG	Sonstige	0.52	0.38	0.66
KG	Schweiz	0.63	0.59	0.66
KG	Europa ex Schweiz	0.53	0.45	0.61
KG	Sonstige	0.37	0.23	0.51
HMS	Schweiz	-0.09	-0.23	0.04
HMS	Europa ex Schweiz	-0.53	-0.91	-0.15
HMS	Sonstige	-1.14	-1.88	-0.38
FMS	Schweiz	0.56	0.48	0.64
FMS	Europa ex Schweiz	0.55	0.35	0.75
FMS	Sonstige	0.63	0.26	0.99
IMS	Schweiz	0.39	0.21	0.58
IMS	Europa ex Schweiz	0.37	-0.01	0.76
IMS	Sonstige	-0.25	-1.08	0.58
BM1	Schweiz	-0.14	-0.20	-0.08
BM1	Europa ex Schweiz	-0.31	-0.50	-0.13
BM1	Sonstige	-0.22	-0.50	0.07

Effektstärken des Regimewechsels nach Nationalität auf die Prüfungsnote Deutsch (Gesamt)

Prüfung	Nationalität	Effektstärke d	CI-	CI+
LG	Schweiz	0.57	0.54	0.60
LG	Europa ex Schweiz	0.54	0.47	0.61
LG	Sonstige	0.58	0.44	0.72
KG	Schweiz	0.31	0.28	0.34
KG	Europa ex Schweiz	0.24	0.16	0.32
KG	Sonstige	0.26	0.12	0.40
HMS	Schweiz	0.27	0.14	0.41
HMS	Europa ex Schweiz	-0.17	-0.54	0.20
HMS	Sonstige	0.18	-0.52	0.89
FMS	Schweiz	0.80	0.72	0.89
FMS	Europa ex Schweiz	0.62	0.42	0.82
FMS	Sonstige	0.63	0.26	0.99
IMS	Schweiz	0.12	-0.06	0.31
IMS	Europa ex Schweiz	0.25	-0.14	0.63
IMS	Sonstige	-0.25	-1.08	0.58
BM1	Schweiz	0.39	0.33	0.45
BM1	Europa ex Schweiz	0.15	-0.04	0.33
BM1	Sonstige	0.21	-0.07	0.50

Effektstärken des Regimewechsels nach Nationalität auf die Prüfungsnote Deutsch (Aufsatz)

Prüfung	Nationalität	Effektstärke d	CI-	CI+
LG	Schweiz	0.53	0.50	0.56
LG	Europa ex Schweiz	0.51	0.44	0.58
LG	Sonstige	0.63	0.49	0.77
KG	Schweiz	0.51	0.48	0.54
KG	Europa ex Schweiz	0.45	0.37	0.53
KG	Sonstige	0.46	0.32	0.59
HMS	Schweiz	0.56	0.43	0.70
HMS	Europa ex Schweiz	0.19	-0.18	0.56
HMS	Sonstige	0.53	-0.19	1.24
FMS	Schweiz	0.15	0.07	0.23
FMS	Europa ex Schweiz	0.03	-0.16	0.23
FMS	Sonstige	0.07	-0.29	0.42
IMS	Schweiz	0.06	-0.12	0.25
IMS	Europa ex Schweiz	0.16	-0.22	0.54
IMS	Sonstige	-0.39	-1.22	0.45

Effektstärken des Regimewechsels nach Nationalität auf die Prüfungsnote Deutsch (Grammatik)

Prüfung	Nationalität	Effektstärke d	CI-	CI+
LG	Schweiz	0.50	0.47	0.53
LG	Europa ex Schweiz	0.46	0.39	0.53
LG	Sonstige	0.41	0.27	0.55
KG	Schweiz	0.06	0.02	0.09
KG	Europa ex Schweiz	0.01	-0.07	0.08
KG	Sonstige	0.04	-0.10	0.18
HMS	Schweiz	-0.07	-0.21	0.06
HMS	Europa ex Schweiz	-0.40	-0.77	-0.02
HMS	Sonstige	-0.17	-0.87	0.54
FMS	Schweiz	1.28	1.19	1.36
FMS	Europa ex Schweiz	1.08	0.87	1.29
FMS	Sonstige	0.99	0.62	1.36
IMS	Schweiz	0.15	-0.03	0.34
IMS	Europa ex Schweiz	0.28	-0.10	0.66
IMS	Sonstige	-0.01	-0.83	0.82

Tabelle 6

Effektstärken der Notenveränderung nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und nach der Muttersprache der Kandidat*innen

Effektstärken des Regimewechsels nach Muttersprache auf die Prüfungsnote Mathematik

Prüfung	Muttersprache	Effektstärke d	CI-	CI+
LG	DEU	0.54	0.51	0.57
LG	SGR	0.58	0.49	0.67
LG	NGR	0.57	0.47	0.67
LG	ETC	0.52	0.41	0.63
KG	DEU	0.62	0.58	0.65
KG	SGR	0.60	0.51	0.69
KG	NGR	0.58	0.46	0.70
KG	ETC	0.52	0.41	0.62
HMS	DEU	-0.18	-0.32	-0.04
HMS	SGR	-0.08	-0.50	0.33
HMS	NGR	0.22	-0.33	0.76
HMS	ETC	-0.45	-0.97	0.08
FMS	DEU	0.53	0.45	0.61
FMS	SGR	0.79	0.56	1.03
FMS	NGR	0.47	0.14	0.81
FMS	ETC	0.71	0.45	0.96
IMS	DEU	0.31	0.11	0.51
IMS	SGR	0.26	-0.17	0.68
IMS	NGR	1.02	0.38	1.65
IMS	ETC	0.76	0.16	1.36
BM1	DEU	-0.16	-0.22	-0.10
BM1	SGR	-0.14	-0.36	0.08
BM1	NGR	-0.18	-0.38	0.03
BM1	ETC	-0.12	-0.34	0.11

Effektstärken des Regimewechsels nach Muttersprache auf die Prüfungsnote Deutsch (Gesamt)

Prüfung	Muttersprache	Effektstärke d	CI-	CI+
LG	DEU	0.61	0.58	0.64
LG	SGR	0.67	0.58	0.76
LG	NGR	0.61	0.51	0.72
LG	ETC	0.59	0.48	0.70
KG	DEU	0.33	0.30	0.37
KG	SGR	0.44	0.35	0.53
KG	NGR	0.27	0.16	0.39
KG	ETC	0.25	0.15	0.36
HMS	DEU	0.27	0.13	0.41
HMS	SGR	0.03	-0.38	0.44
HMS	NGR	-0.23	-0.78	0.31
HMS	ETC	0.17	-0.35	0.70
FMS	DEU	0.85	0.77	0.94
FMS	SGR	0.79	0.56	1.03
FMS	NGR	0.54	0.20	0.87
FMS	ETC	0.76	0.50	1.01
IMS	DEU	0.05	-0.14	0.25
IMS	SGR	0.26	-0.16	0.69
IMS	NGR	0.51	-0.10	1.11
IMS	ETC	0.35	-0.24	0.93
BM1	DEU	0.38	0.32	0.45
BM1	SGR	0.30	0.08	0.52
BM1	NGR	0.16	-0.04	0.37
BM1	ETC	0.34	0.11	0.57

Effektstärken des Regimewechsels nach Muttersprache auf die Prüfungsnote Deutsch (Aufsatz)

Prüfung	Muttersprache	Effektstärke d	CI-	CI+
LG	DEU	0.54	0.51	0.57
LG	SGR	0.66	0.57	0.75
LG	NGR	0.61	0.51	0.72
LG	ETC	0.67	0.56	0.78
KG	DEU	0.54	0.51	0.58
KG	SGR	0.55	0.46	0.63
KG	NGR	0.46	0.34	0.57
KG	ETC	0.46	0.36	0.57
HMS	DEU	0.60	0.46	0.75
HMS	SGR	0.22	-0.20	0.63
HMS	NGR	-0.08	-0.62	0.47
HMS	ETC	0.47	-0.06	0.99
FMS	DEU	0.18	0.09	0.26
FMS	SGR	0.03	-0.20	0.26
FMS	NGR	-0.01	-0.34	0.32
FMS	ETC	0.26	0.01	0.51
IMS	DEU	0.01	-0.18	0.21
IMS	SGR	0.05	-0.38	0.47
IMS	NGR	0.34	-0.27	0.94
IMS	ETC	0.31	-0.28	0.89

Effektstärken des Regimewechsels nach Muttersprache auf die Prüfungsnote Deutsch (Grammatik)

Prüfung	Muttersprache	Effektstärke d	CI-	CI+
LG	DEU	0.55	0.51	0.58
LG	SGR	0.55	0.46	0.64
LG	NGR	0.49	0.39	0.59
LG	ETC	0.40	0.29	0.51
KG	DEU	0.06	0.02	0.09
KG	SGR	0.25	0.16	0.33
KG	NGR	0.05	-0.06	0.17
KG	ETC	0.03	-0.08	0.13
HMS	DEU	-0.10	-0.24	0.04
HMS	SGR	-0.13	-0.55	0.28
HMS	NGR	-0.28	-0.83	0.27
HMS	ETC	-0.14	-0.66	0.38
FMS	DEU	1.32	1.23	1.41
FMS	SGR	1.37	1.11	1.62
FMS	NGR	0.94	0.59	1.28
FMS	ETC	1.00	0.74	1.26
IMS	DEU	0.08	-0.11	0.28
IMS	SGR	0.41	-0.02	0.84
IMS	NGR	0.59	-0.02	1.20
IMS	ETC	0.28	-0.30	0.86

Tabelle 7

Effektstärken der Notenveränderung nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und nach der Aufenthaltsdauer der Kandidat*innen

Effektstärken des Regimewechsels nach Aufenthaltsdauer auf die Prüfungsnote Mathematik (nur 2022/2023)

Prüfung	Aufenthaltsdauer	Effektstärke d	CI-	CI+
LG	bis zu 10 Jahre	0.43	0.29	0.56
LG	mehr als 10 Jahre	0.56	0.44	0.68
LG	einheimisch	0.59	0.55	0.63
KG	bis zu 10 Jahre	0.71	0.57	0.85
KG	mehr als 10 Jahre	0.76	0.62	0.89
KG	einheimisch	0.76	0.72	0.81
HMS	bis zu 10 Jahre	-0.19	-0.90	0.51
HMS	mehr als 10 Jahre	0.17	-0.48	0.82
HMS	einheimisch	-0.02	-0.20	0.16
FMS	bis zu 10 Jahre	1.16	0.78	1.54
FMS	mehr als 10 Jahre	0.86	0.53	1.19
FMS	einheimisch	0.46	0.37	0.56
IMS	bis zu 10 Jahre	0.41	-0.11	0.94
IMS	mehr als 10 Jahre	0.32	-0.10	0.74
IMS	einheimisch	0.54	0.28	0.79
BM1	bis zu 10 Jahre	-0.29	-0.54	-0.05
BM1	mehr als 10 Jahre	-0.12	-0.31	0.07
BM1	einheimisch	-0.15	-0.23	-0.08

Effektstärken des Regimewechsels nach Aufenthaltsdauer auf die Prüfungsnote Deutsch (Gesamt) (nur 2022/2023)

Prüfung	Aufenthaltsdauer	Effektstärke d	CI-	CI+
LG	bis zu 10 Jahre	0.52	0.39	0.66
LG	mehr als 10 Jahre	0.71	0.59	0.83
LG	einheimisch	0.73	0.69	0.77
KG	bis zu 10 Jahre	0.43	0.29	0.57
KG	mehr als 10 Jahre	0.44	0.31	0.57
KG	einheimisch	0.46	0.41	0.50
HMS	bis zu 10 Jahre	-0.13	-0.83	0.58
HMS	mehr als 10 Jahre	0.95	0.25	1.63
HMS	einheimisch	0.40	0.22	0.58
FMS	bis zu 10 Jahre	0.37	0.01	0.73
FMS	mehr als 10 Jahre	0.43	0.11	0.75
FMS	einheimisch	0.99	0.89	1.09
IMS	bis zu 10 Jahre	-0.38	-0.90	0.15
IMS	mehr als 10 Jahre	-0.29	-0.71	0.13
IMS	einheimisch	0.26	0.01	0.52
BM1	bis zu 10 Jahre	0.43	0.19	0.68
BM1	mehr als 10 Jahre	0.38	0.19	0.57
BM1	einheimisch	0.46	0.38	0.53

Effektstärken des Regimewechsels nach Aufenthaltsdauer auf die Prüfungsnote Deutsch (Aufsatz) (nur 2022/2023)

Prüfung	Aufenthaltsdauer	Effektstärke d	CI-	CI+
LG	bis zu 10 Jahre	0.66	0.52	0.80
LG	mehr als 10 Jahre	0.79	0.66	0.91
LG	einheimisch	0.73	0.69	0.77
KG	bis zu 10 Jahre	0.51	0.37	0.64
KG	mehr als 10 Jahre	0.53	0.40	0.66
KG	einheimisch	0.63	0.59	0.68
HMS	bis zu 10 Jahre	-0.15	-0.85	0.56
HMS	mehr als 10 Jahre	0.86	0.17	1.53
HMS	einheimisch	0.74	0.56	0.92
FMS	bis zu 10 Jahre	0.01	-0.35	0.37
FMS	mehr als 10 Jahre	0.08	-0.24	0.39
FMS	einheimisch	0.29	0.19	0.38
IMS	bis zu 10 Jahre	-0.69	-1.22	-0.15
IMS	mehr als 10 Jahre	-0.35	-0.77	0.07
IMS	einheimisch	0.03	-0.23	0.28

Effektstärken des Regimewechsels nach Aufenthaltsdauer auf die Prüfungsnote Deutsch (Grammatik) (nur 2022/2023)

Prüfung	Aufenthaltsdauer	Effektstärke d	CI-	CI+
LG	bis zu 10 Jahre	0.32	0.18	0.45
LG	mehr als 10 Jahre	0.50	0.38	0.62
LG	einheimisch	0.57	0.53	0.61
KG	bis zu 10 Jahre	0.27	0.13	0.41
KG	mehr als 10 Jahre	0.24	0.11	0.37
KG	einheimisch	0.18	0.14	0.23
HMS	bis zu 10 Jahre	-0.08	-0.79	0.62
HMS	mehr als 10 Jahre	0.60	-0.07	1.26
HMS	einheimisch	-0.01	-0.19	0.17
FMS	bis zu 10 Jahre	0.66	0.30	1.03
FMS	mehr als 10 Jahre	0.70	0.38	1.03
FMS	einheimisch	1.43	1.33	1.54
IMS	bis zu 10 Jahre	0.01	-0.52	0.53
IMS	mehr als 10 Jahre	-0.12	-0.53	0.30
IMS	einheimisch	0.45	0.19	0.70

Tabelle 8

Effektstärken der Notenveränderung nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und nach der Trägerschaft der abgebenden Schule der Kandidat*innen

Effektstärken des Regimewechsels nach Trägerschaft des abgebenden Schultyps auf die Prüfungsnote Mathematik (nur 2022/2023)

Prüfung	Trägerschaft	Effektstärke d	CI-	CI+
LG	öffentlich	0.53	0.50	0.57
LG	privat	0.54	0.36	0.72
KG	öffentlich	0.57	0.53	0.62
KG	privat	0.66	0.54	0.78
HMS	öffentlich	-0.31	-0.53	-0.08
HMS	privat	-0.20	-0.77	0.37
FMS	öffentlich	0.83	0.70	0.96
FMS	privat	0.85	0.52	1.19
IMS	öffentlich	0.36	0.15	0.56
IMS	privat	0.46	-0.06	0.99
BM1	öffentlich	-0.16	-0.22	-0.10
BM1	privat	-0.16	-0.38	0.05

Effektstärken des Regimewechsels nach Trägerschaft des abgebenden Schultyps auf die Prüfungsnote Deutsch (Gesamt) (nur 2022/2023)

Prüfung	Trägerschaft	Effektstärke d	CI-	CI+
LG	öffentlich	0.53	0.49	0.57
LG	privat	0.60	0.42	0.78
KG	öffentlich	0.27	0.22	0.31
KG	privat	0.32	0.20	0.44
HMS	öffentlich	0.31	0.09	0.53
HMS	privat	0.17	-0.40	0.74
FMS	öffentlich	0.33	0.20	0.46
FMS	privat	0.29	-0.03	0.62
IMS	öffentlich	0.20	-0.01	0.40
IMS	privat	0.25	-0.27	0.77
BM1	öffentlich	0.37	0.31	0.43
BM1	privat	0.13	-0.08	0.35

Effektstärken des Regimewechsels nach Trägerschaft des abgebenden Schultyps auf die Prüfungsnote Deutsch (Aufsatz) (nur 2022/2023)

Prüfung	Trägerschaft	Effektstärke d	CI-	CI+
LG	öffentlich	0.53	0.49	0.57
LG	privat	0.57	0.38	0.75
KG	öffentlich	0.50	0.45	0.54
KG	privat	0.50	0.39	0.62
HMS	öffentlich	0.37	0.14	0.59
HMS	privat	0.27	-0.31	0.84
FMS	öffentlich	0.11	-0.02	0.24
FMS	privat	-0.05	-0.38	0.27
IMS	öffentlich	0.17	-0.03	0.38
IMS	privat	0.01	-0.51	0.53

Effektstärken des Regimewechsels nach Trägerschaft des abgebenden Schultyps auf die Prüfungsnote Deutsch (Grammatik) (nur 2022/2023)

Prüfung	Trägerschaft	Effektstärke d	CI-	CI+
LG	öffentlich	0.43	0.40	0.47
LG	privat	0.55	0.36	0.73
KG	öffentlich	0.02	-0.02	0.07
KG	privat	0.09	-0.03	0.21
HMS	öffentlich	0.19	-0.04	0.41
HMS	privat	0.03	-0.54	0.60
FMS	öffentlich	0.49	0.36	0.62
FMS	privat	0.63	0.30	0.96
IMS	öffentlich	0.16	-0.05	0.36
IMS	privat	0.45	-0.08	0.98

Hinsichtlich der Aufenthaltsdauer in der Schweiz zeigten sich die meisten Gruppenunterschiede, wie in **Tabelle** ersichtlich, wobei noch einmal betont werden soll, dass für die meisten Prüfungsnoten und Prüfungstypen keine Unterschiede nachgewiesen werden konnten. Bei der Mathematikprüfung der FMS haben Jugendliche mit einer Aufenthaltsdauer von bis zu zehn Jahren (d = 1.16; CI 0.53–1.19) mit dem Regimewechsel ihre Noten etwas mehr verbessert als einheimische Jugendliche (d = 0.46; CI 0.37–0.56), wobei die mittlere Effektstärke des Regimewechsels bei d = .48 lag.

Genau umgekehrt stellte sich das bei der Prüfungsnote Deutsch (Gesamt) bei der gleichen FMS-Prüfung dar. Hier haben einheimische Jugendliche (*d* = .99; CI 0.93–1.05) nach dem Regimewechsel ihre Noten etwas mehr verbessert als Jugendliche mit einer Aufenthaltsdauer von bis zu zehn Jahren (*d* = .37; CI 0.30–0.44), wobei die durchschnittliche Differenz infolge des Regimewechsels hier bei *d* = 0.79 lag. Der letztere Gruppenunterschied ist in erster Linie nicht auf die Prüfungsnote Deutsch (Aufsatz), sondern auf die Prüfungsnote Deutsch (Grammatik) zurückzuführen. Hier haben einheimische Jugendliche (*d* = 1.43; CI 1.36–1.50) eher profitiert als Jugendliche mit einer Aufenthaltsdauer von bis zu zehn Jahren (*d* = .66; CI 0.58–0.74), wobei die durchschnittliche Effektstärke des Regimewechsels hier bei *d* = 1.29 lag. Für die Prüfungsnote Deutsch (Aufsatz) zeigte sich dieser Gruppenunterschied nicht. Für die Prüfungsnote Deutsch (Grammatik) zeigte sich für die LG-Prüfung ausserdem, dass einheimische Jugendliche (*d* = .57; CI 0.49–0.65) eher profitiert haben als Jugendliche mit einer Aufenthaltsdauer von bis zu zehn Jahren (*d* = .32; CI 0.24–0.40).

Schliesslich wird aus **Tabelle** deutlich, dass sich bezüglich der Trägerschaft der abgebenden Schule keine auffälligen Gruppenunterschiede zeigten. Mit anderen Worten sind zwar die Prüfungsnoten insgesamt gestiegen, aber sie taten dies nicht in einem unterschiedlichen Ausmass für Abgänger*innen von öffentlichen und für Abgänger*innen von privaten Schulen.

Über alle betrachteten Gruppenvariablen hinweg zeigt sich insgesamt ein konsistentes Muster: Die Notenverbesserungen post-VAM (2023-2024) gegenüber ante-VAM (2020-2022) über die untersuchten Gruppen hinweg sehr ähnlich aus. Das wird insbesondere daran sichtbar, dass sich die Konfidenzintervalle zwischen den Gruppen meist überlappen und die geschätzten standardisierten Differenzen innerhalb eines Prüfungstyps typischerweise nah beieinanderliegen. Dieses Ergebnis spricht dafür, dass der Regimevergleich (2020-2022 vs. 2023-2024) in Bezug auf die Notenveränderungen nur selten mit gruppenspezifischen Differenzen verbunden ist.

Die wenigen Ausnahmen lassen sich präzise benennen und sollten entsprechend als «lokale» Befunde verstanden werden: Beim Geschlecht treten zwei Abweichungen in der Teilnote Deutsch (Aufsatz) auf (KG: stärkere Verbesserung bei Knaben als bei Mädchen; IMS: umgekehrtes Muster, das jedoch auf sehr wenige leistungsstarke Mädchen zurückgeführt wird). Bei der Nationalität findet sich als Ausnahme die Mathematiknote in der KG-Prüfung, bei der Kandidat*innen mit Schweizer Nationalität eine etwas stärkere Verbesserung zeigen als Kandidat*innen mit sonstiger Nationalität. Die meisten gruppenspezifischen Unterschiede treten bei der Aufenthaltsdauer auf, insbesondere in der FMS, wobei sich die Richtung je nach Fach unterscheidet (beispielsweise stärkere Verbesserung in Mathematik bei «bis zu 10 Jahre», aber stärkere Verbesserung in Deutsch (Gesamt/Grammatik) bei «einheimisch»). Für die Trägerschaft der abgebenden Schule zeigen sich bei den Notenveränderungen dagegen keine auffälligen Unterschiede. Diese Ergebnisse sprechen dafür, dass der Regimevergleich (2020-2022 vs. 2023-2024) in Bezug

auf die Notenveränderungen nicht mit systematischen Veränderungen zwischen den Gruppen einhergeht

Für die Einordnung dieser Ausnahmen sind zwei methodische Punkte wichtig. Erstens wird hier eine grosse Zahl von Vergleichen berichtet (mehrere Fächer/Teilnoten × mehrere Prüfungstypen × mehrere Gruppen). Dadurch ist es erwartbar, dass einzelne Abweichungen auftreten, ohne dass dahinter zwingend ein systematisches Muster stehen muss. Entsprechend ist es sinnvoll, vor allem auf Befunde Gewicht zu legen, die innerhalb eines Prüfungstyps über mehrere Prüfungsfächer oder mehrere Teilbereiche hinweg konsistent sind oder sich in mehreren Prüfungstypen in ähnlicher Richtung zeigen (Benjamini & Hochberg, 1995). Solche Verschiebungen zeigen sich in den vorliegenden Daten nicht. Zweitens können einzelne Teilgruppen – je nach Gruppenvariable und Prüfungstyp – klein sein, wodurch Effektstärken stärker durch wenige Fälle beeinflusst werden können (wie in diesem Abschnitt zur IMS-Ausnahme explizit beschrieben). Insgesamt bleibt damit als Kernaussage festzuhalten: Die Notenverbesserungen im Regimevergleich sind breit sichtbar, aber in Bezug auf die untersuchten Gruppenvariablen nur selten differenziell ausgeprägt; dort, wo Differenzen auftreten, sollten sie als gezielt zu prüfende Hinweise verstanden werden, nicht als generelle oder systematische Gruppenmuster.

4. Zusammenhang zwischen Vornote und Prüfungsnote

Bei der folgenden Fragestellung ging es darum, herauszufinden, wie hoch die Vornoten mit den Prüfungsnoten korreliert sind. Grundsätzlich würde man eine positive Korrelation erwarten, deren Grössenordnung etwas darüber aussagt, wie redundant beide Masse sind. Eine *hohe* positive Korrelation würde dafürsprechen, dass die Vornoten und die Prüfungsnoten im Wesentlichen das Gleiche messen, sodass man auf eine davon verzichten könnte, ohne viel Information zu verlieren. Man würde so beispielsweise einen Befund erwarten, wenn man davon ausginge, dass beide Noten ausschliesslich von den kognitiven Fähigkeiten abhängen, die über die Zeit weitgehend stabil sind. Eine *niedrige* positive Korrelation würde dafürsprechen, dass mit den Vornoten und den Prüfungsnoten unterschiedliche Fähigkeiten gemessen werden. Es könnte beispielsweise sein, dass die Vornoten eher so etwas wie die Dauerleistungsfähigkeit messen, während die Prüfungsnoten die Spitzenleistungsfähigkeit abbilden. Zu beachten ist bei diesen Überlegungen, dass damit nicht die Frage beantwortet wird, welche der beiden Noten eine höhere Vorhersagekraft für den zukünftigen Lernerfolg hat. Es wird lediglich untersucht, in welchem Ausmass das gleiche Merkmal gemessen wird.

Die berichteten Korrelationen zwischen Vornoten und Prüfungsnoten liegen insgesamt im Bereich von $.40 < r < .60$ und damit in einer Grössenordnung, die auf eine teilweise, aber keineswegs vollständige Überlappung der beiden Leistungsindikatoren hinweist.

Beispielsweise liegt die Korrelation der Mathematiknoten für das Langgymnasium ante-VAM (2020-2022) bei $r = .52$ und post-VAM (2023-2024) bei $r = .53$ bzw. die Korrelation der Deutschnoten ante-VAM (2020-2022) bei $r = .58$ und post-VAM (2023-2024) bei $r = .58$ und damit im Bereich mittlerer Effektstärke. Vornote und Prüfungsnote erfassen demnach gemeinsame Anteile schulischer Leistungsfähigkeit, enthalten aber jeweils auch spezifische Informationen. Das passt auch zur inhaltlichen Differenzierung, dass Vornoten eher Leistungsentwicklungen und kontinuierliche Leistungsanforderungen im Unterricht abbilden können, während Prüfungsnoten stärker die Leistung in einer standardisierten, zeitlich verdichteten Prüfungssituation widerspiegeln.

Der Zusammenhang für das LG fällt höher ($r > .50$) aus als für die übrigen Prüfungstypen ($r < .50$). In der Logik der hier verfolgten Fragestellung bedeutet dies, dass Vornote und Prüfungsnote gerade ausserhalb des LG weniger redundant sind und damit tendenziell komplementäre Hinweise zur Leistungsfähigkeit liefern. Für das LG liegen Vornote und Prüfungsnote sowohl ante-VAM (2020-2022) als auch post-VAM (2023-2024) vor; der Zusammenhang bleibt dabei durch den Regimewechsel praktisch unverändert. Bei der Interpretation der Höhe einer Korrelation ist generell mitzudenken, dass weder sehr niedrige noch sehr hohe Werte automatisch «gute» oder «schlechte» Messungen bedeuten. Korrelationen hängen unter anderem davon ab, wie breit die Leistungsstreuung in der betrachteten Population ist und in welchem Ausmass beide Indikatoren von situativen und kontextuellen Einflüssen geprägt sind (Schober et al., 2018). Eine nahezu vollständige Redundanz wäre erst bei sehr hohen Zusammenhängen zu erwarten (etwa $r = .80$ oder $r = .90$); davon ist hier jedoch klar nicht auszugehen. Insgesamt stützen die Befunde somit die Schlussfolgerung, dass die Vornote neben der Prüfungsnote zusätzliche Information über die Leistungsfähigkeit der Kandidat*innen liefert. Ob und in welchem Ausmass diese zusätzliche Information auch für die Vorhersage weiterer Kriterien (etwa im späteren Verlauf) bedeutsam ist, wird in den folgenden Analysen gesondert betrachtet.

5. Vornoten und deren Streuung zwischen Einheiten

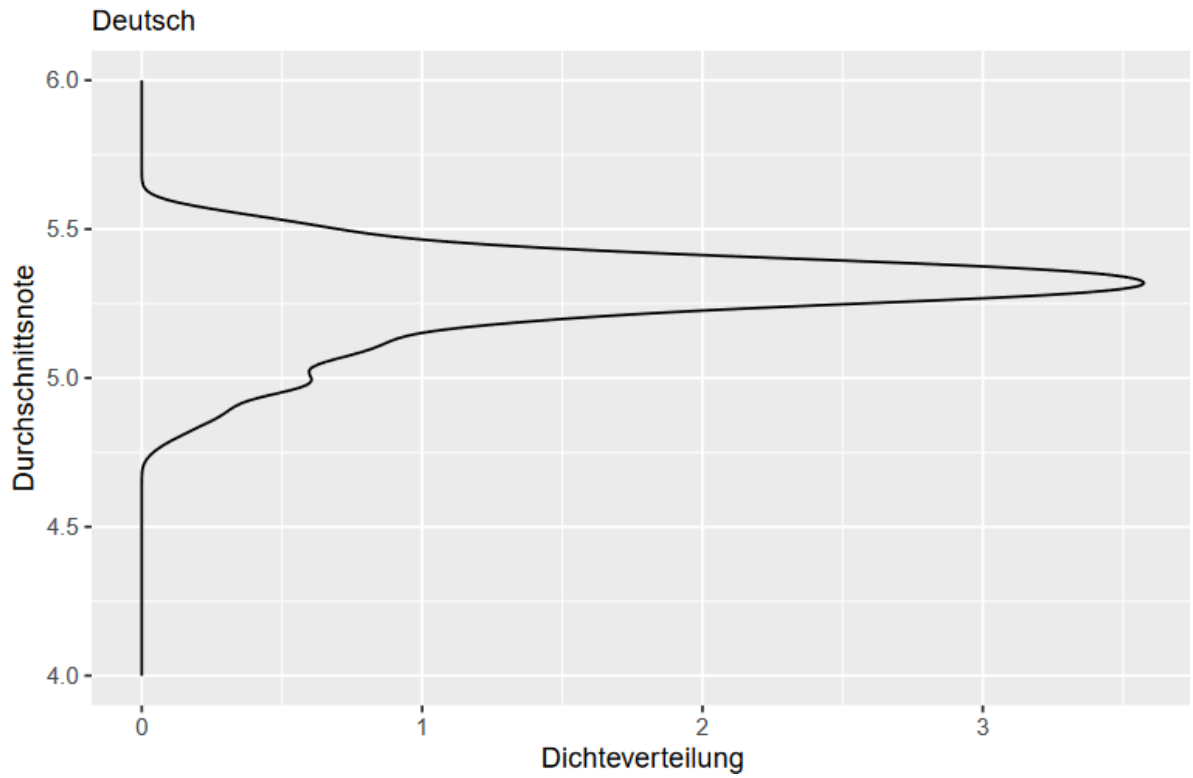
Weil unter dem neuen Prüfungsregime die Vornoten eine wesentliche Rolle spielen, ist es informativ zu erfahren, wie diese zwischen den einzelnen Schulen und Schulgemeinden streuen. Man kann rechnerisch die gesamte Varianz in den Vornoten (also deren Streuung über alle Jugendlichen hinweg) den einzelnen Organisationsebenen (also Jugendlichen, Schulen, Schulgemeinden und Schulbezirken) zuordnen und die Ähnlichkeit der Vornoten auf den höheren Organisationsebenen (also Schulen, Schulgemeinden und Schulbezirken) mit der Intra-Class-Korrelation (ICC) ausdrücken. Diese kann Werte zwischen 0 und 1 annehmen, wobei eine ICC = 0 bedeuten würde, dass man aus dem Wissen von der Zugehörigkeit zu einer höheren Organisationseinheit keinerlei Schlüsse auf die Vornote einer oder eines einzelnen Jugendlichen ziehen kann. Bei einer ICC = 1 dagegen könnte man aus dem Wissen zur Zugehörigkeit zu einer höheren

Organisationseinheit die Vornote einer oder eines einzelnen Jugendlichen perfekt vorhersagen, weil sich die Vornoten innerhalb der höheren Organisationseinheiten nicht unterscheiden. In Bezug auf schulische Leistungen allgemein findet man zwischen Schulen grössenordnungsmässig ICC von .10 < ICC < .30 (siehe Bosker & Witziers, 1996; Hedges & Hedberg, 2007; Stockford, 2009). Zwei Schüler*innen einer bestimmten Schule sind sich in ihren Leistungen also ein wenig ähnlicher als zwei zufällig aus der Population gezogenen Schüler*innen, aber diese Ähnlichkeit ist nicht sehr gross ausgeprägt.

Auswertungen mit den Zürcher Daten zeigen, dass die Vornoten für Deutsch (ICC = .17) und Englisch (ICC = .11) relativ hoch und bei Französisch (ICC = .07) oder Mathematik (ICC = .04) relativ niedrig sind. Mit «relativ» ist hier der Vergleich der vier Schulfächer gemeint, denn von der absoluten Grössenordnung her sind alle diese Zahlen niedrig und bewegen sich im unteren Bereich der Spanne, den man aus anderen Untersuchungen kennt. Ähnliches gilt für die Streuung zwischen den abgebenden Schulgemeinden, wenn man diese Streuung alleine für sich betrachtet. Dagegen ist die Varianzaufklärung durch den Schulbezirk praktisch zu vernachlässigen (ICC < .01) und wird deswegen hier nicht weiter betrachtet.

Abbildung 2

Verteilung der durchschnittlichen Vornoten der abgebenden Schulen im Kanton Zürich



In sogenannten Mehrebenenanalysen zur Varianzdekomposition kann man nun versuchen, die Varianz zwischen Schulen und zwischen Schulgemeinden aufzuteilen. Hierbei zeigt sich, dass es in allererster Linie die Varianz zwischen den Schulen ist, die für

eine hohe ICC sorgt, und die Varianz zwischen Schulgemeinden zu vernachlässigen ist, wenn man für die abgebende Schule kontrolliert.

Die konkrete Schulzugehörigkeit ist also mit eher kleinen, aber dennoch messbaren Unterschieden in der Vornote verbunden, vor allem in Deutsch und Englisch und deutlich weniger in Französisch und Mathematik. In welchem Schulbezirk diese Schule aber angesiedelt ist, spielt dagegen keine Rolle. Mit anderen Worten gibt es in jedem Schulbezirk Schulen, die eher höhere Vornoten vergeben, und solche, die eher niedrigere Vornoten vergeben. **Abbildung 2** zeigt beispielhaft auf, wie stark die Vornoten in Deutsch zwischen Schulen im ganzen Kanton Zürich variieren. Für Deutsch ist die ICC am höchsten und es gibt Schulen, bei denen die Jugendlichen im Schnitt eine Vornote von 4.75 haben, und solche, bei denen die Jugendlichen im Schnitt eine Vornote von über 5.50 mitbringen.

Die in diesem Abschnitt berichteten ICCs lassen sich als Hinweis lesen, in welchem Ausmass Vornoten neben individuellen Leistungsinformationen auch einen schulbezogenen Anteil enthalten: Je höher die ICC, desto ähnlicher sind Vornoten innerhalb derselben Schule und desto grösser ist der Anteil der Varianz, der zwischen Schulen liegt (McGraw & Wong, 1996). Für die Interpretation ist dabei zentral, dass dieser Befund zunächst deskriptiv ist: Unterschiede zwischen Schulen können sowohl mit Unterschieden in der Zusammensetzung der Schüler*innenschaft als auch mit Unterschieden in Beurteilungspraktiken einhergehen, ohne dass sich daraus bereits eine eindeutige Ursache ableiten lässt. Im vorliegenden Kontext unterstützt die Varianzstruktur damit vor allem eine pragmatische Lesart: Vornoten sind als Entscheidungskomponenten nicht rein «individuell» im Sinne eines vollständig schulinvarianten Signals, sondern enthalten – je nach Fach – in unterschiedlichem Ausmass auch schulgebundene Information. Die Grössenordnung bewegt sich dabei im Rahmen dessen, was in schulischen Leistungsdaten häufig beobachtet wird (Hedges & Hedberg, 2007).

6. Vornoten von Jugendlichen, die bestanden haben oder die abgelehnt wurden

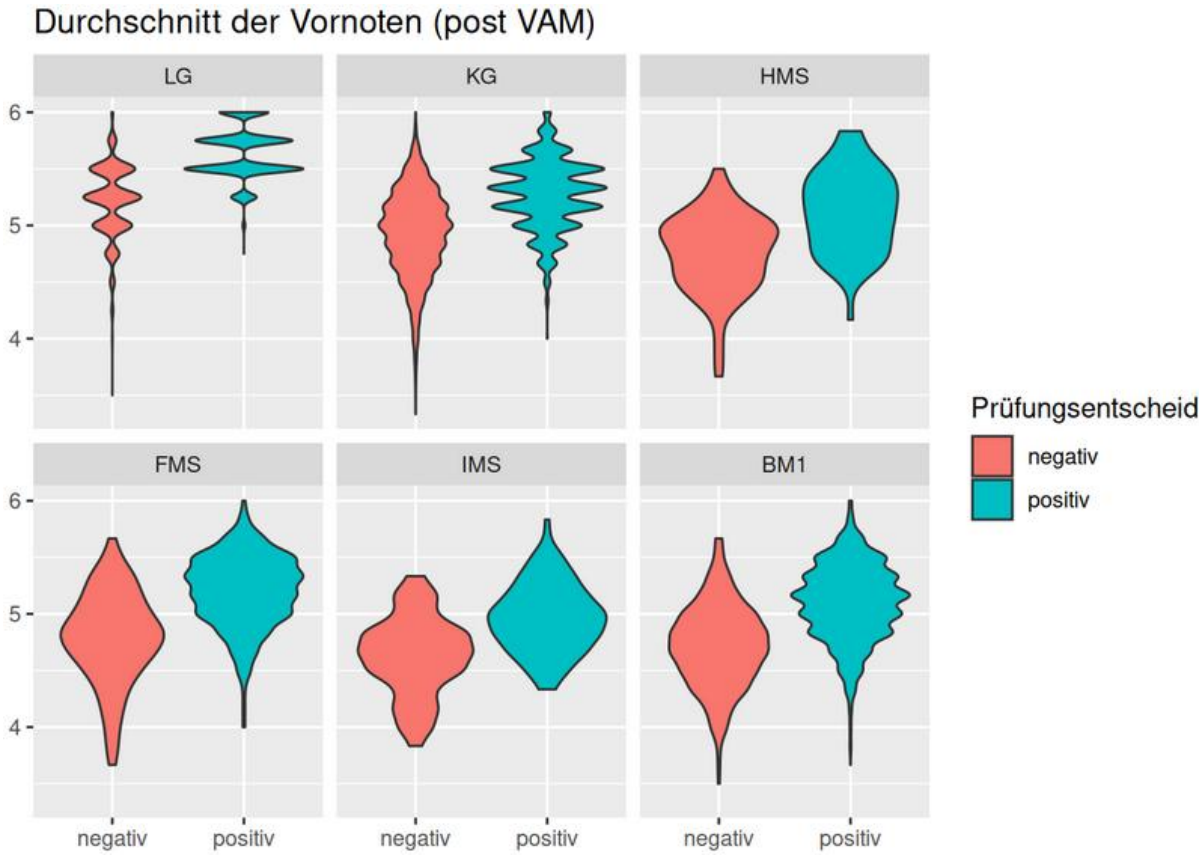
Ergänzend zu den korrelativen Auswertungen zum Zusammenhang zwischen der Vornote und der Prüfungsnote war es von Interesse, wie stark sich die Vornoten von Jugendlichen unterscheiden, die bestanden haben oder abgelehnt wurden. Man würde hier grundsätzlich erwarten, dass die Vornoten von Jugendlichen, die bestanden haben, im Durchschnitt höher sind als die Vornoten von Jugendlichen, die abgelehnt wurden. Gleichzeitig würde man aber von einer gewissen Überlappung ausgehen, weil man eine schwächere Vornote mit einer höheren Prüfungsnote kompensieren kann und umgekehrt.

Aus **Abbildung 3** ist ersichtlich, dass die Vornoten von Jugendlichen, die bestanden haben, höher sind als die Vornoten von Jugendlichen, die abgelehnt wurden. Es zeigt sich

aber auch eine hohe Überlappung der Vornoten von Jugendlichen, die bestanden haben, und solchen, die abgelehnt wurden. Man kann also mit eher schwächeren Vornoten bestehen und mit eher starken Vornoten abgelehnt werden. Dieser Befund ist im Einklang mit den in Abschnitt 4 berichteten geringen Korrelationen zwischen Vornoten und Prüfungsnoten.

Abbildung 3

Vornotendurchschnitt nach dem Regimewechsel, aufgeteilt nach Prüfungsentscheid und den einzelnen Prüfungstypen



Konkret zeigt sich beispielsweise, dass es bei der LG-Prüfung Fälle gibt, bei denen Jugendliche mit einer relativ hohen Vornote von beispielsweise 5.50 abgelehnt werden, oder dass es bei der KG-Prüfung zumindest einzelne Jugendliche gibt, die mit einer Vornote von beispielsweise nur 4.50 trotzdem bestehen können. Des Weiteren fällt auf, dass es bei einzelnen Prüfungstypen wie der HMS- oder der IMS-Prüfung keine Jugendlichen gibt, die sich mit sehr hohen Vornoten (also etwa über 5.75) für die Prüfung anmelden. Auch lässt sich aus dem Befund die Beobachtung ableiten, dass es gerade bei den Mittelschulprüfungen über einer bestimmten Vornote, die beispielsweise für die HMS-Prüfung bei etwa 5.20 liegt, praktisch keine negativen Prüfungsentscheide gibt.

Die dargestellten Verteilungen machen sichtbar, dass die Vornoten zwischen Jugendlichen, die bestanden haben, und solchen, die abgelehnt wurden, sich zwar im

Mittel unterscheiden, aber keine scharfe Trennlinie bilden. Praktisch heisst das: Im mittleren Bereich der Vornoten ist die Entscheidung nicht durch die Vornote «deterministisch», sondern entsteht aus der Kombination beider Komponenten; entsprechend ist eine deutliche Überlappung erwartbar und konsistent mit den zuvor berichteten eher moderaten Zusammenhängen zwischen Vornote und Prüfungsnote. Gleichzeitig liefern die Randbereiche eine wichtige Verständnishilfe: Dort, wo sehr hohe Vornoten in einzelnen Prüfungstypen kaum vorkommen oder wo oberhalb einer bestimmten Vornote praktisch keine negativen Entscheide beobachtet werden, ist bei der Interpretation auch an Selbstselektion in die jeweilige Prüfung und an kleine Fallzahlen zu denken. Wenn man die Trennschärfe der Vornote formal ausdrücken wollte, wären dafür Kennwerte aus der Klassifikationsdiagnostik (beispielsweise ROC/AUC) geeignet; die Abbildung liefert dafür bereits die intuitive Grundlage, ohne dass daraus allein eine kausale Interpretation abgeleitet werden sollte (Fawcett, 2006).

7. Kontrafaktisches Szenario

Durch die Berücksichtigung von Vornoten im neuen Prüfungsregime ergibt sich für die Kandidat*innen die Möglichkeit, die Aufnahmeprüfung zu bestehen, auch wenn das Prüfungsergebnis an sich nicht ausreichend gewesen wäre. Das ist immer dann der Fall, wenn die Vornoten hoch genug sind, um das Prüfungsergebnis auszugleichen. Gleichzeitig gibt es auch die prinzipielle Möglichkeit, dass Jugendliche mit einem ausreichenden Prüfungsergebnis die Prüfung nicht bestehen, wenn ihre Vornoten zu niedrig sind. Um zu untersuchen, ob und, wenn ja, wie viele Jugendliche von diesen beiden Situationen betroffen sind, wurde kontrafaktisch untersucht, wie sich die Zulassungszahlen darstellen würden, wenn die Vornoten nicht zählten.

Bei der Interpretation der Ergebnisse in diesem kontrafaktischen Szenario muss beachtet werden, dass es auf Annahmen beruht, die eigentlich nicht realistisch sind. So wird implizit davon ausgegangen, dass, auch wenn die Vornoten nicht zählten, sich am Inhalt der Prüfung, an der vorgenommenen Skalierung, an den vorher festgelegten Bestehensgrenzen, an der Herangehensweise der Jugendlichen an die Prüfung, an ihrer Motivation oder Prüfungsängstlichkeit usw. nichts geändert hätte. Davon ist wahrscheinlich nicht auszugehen. Trotzdem kann man die Ergebnisse zumindest als Hinweis dafür sehen, welchen Einfluss das neue Prüfungsregime mit der Berücksichtigung der Vornoten auf das Bestehen oder Nichtbestehen der Prüfung haben könnte. Allerdings sollten die Ergebnisse dieses Szenarios eher als ein theoretisches Modell verstanden werden, das zur weiteren Diskussion anregt, aber keine endgültigen Schlussfolgerungen über die tatsächlichen Auswirkungen von Vornoten auf das Prüfungsergebnis zulässt.

Die Ergebnisse des kontrafaktischen Szenarios sind für alle Prüfungstypen sehr ähnlich, sodass an dieser Stelle jene für die LG- und die KG-Prüfung herausgegriffen werden. Sie finden sich sowohl in **Tabelle 1** als auch in **Abbildung 4**. Beide enthalten die gleiche

Information und sind auch von der Struktur her gleich aufgebaut. Für das Verständnis der Abbildung bzw. der Tabelle ist zentral, dass hier zwei Entscheidungsregime gegenübergestellt werden: das tatsächliche Regime (inklusive Vornote) und ein rein hypothetisches Regime (ohne Vornote). Jeder der Kandidierenden fällt damit in einen von vier Fällen. Im rechten oberen Quadranten finden sich jene, die unter beiden Regimen bestanden hätten und im linken unteren Quadranten jene, die unter beiden Regimen nicht bestanden hätten. Für diese Kandidierenden macht also der Regimewechsel keinen Unterschied. Aufschlussreich sind die «Wechsel»-Felder, weil dort der Regimewechsel einen Unterschied für das Bestehen ausmacht. Im rechten unteren Quadranten finden sich jene Kandidierende, die nur unter dem neuen aber nicht unter dem alten Regime bestanden hätten. Sie profitieren also von der Einführung der Berücksichtigung der Vornote, weil sie damit ein allein nicht ausreichendes Prüfungsergebnis kompensieren können. Im linken oberen Quadranten finden sich dagegen jene Kandidierende, die unter dem neuen Regime nicht bestanden haben, unter dem alten Regime jedoch bestanden hätten. Deren Vornoten waren so schwach, dass sie ein an sich ausreichendes Prüfungsergebnis abgewertet haben. Die Netto-Differenz zwischen den beiden letztgenannten Feldern beschreibt dabei nicht die «Wirkung» der Vornote im kausalen Sinn, sondern lediglich die rechnerische Verschiebung innerhalb der getroffenen Annahmen.

Tabelle 1

Kontrafaktische Prüfungstestung und faktische Prüfungsleistung des LG- und KG- Prüfungstyps nach faktischem Prüfungsentscheid

Prüfung	kontrafaktisch	nicht bestanden	bestanden
LG	bestanden		965
LG	nicht bestanden	2089	1433
KG	bestanden	10	1142
KG	nicht bestanden	1984	464

Sowohl im rechten oberen Quadranten der Tabelle als auch im rechten oberen Quadranten der Abbildung finden sich also jene Jugendliche, die sowohl nach dem neuen Regime post-VAM (also faktisch) bestanden haben und auch bestanden hätten, wenn das alte Regime ante-VAM (also kontrafaktisch) gegolten hätte. Von den insgesamt 4’487 LG-Prüfungen bzw. 3’600 KG-Prüfungen wäre das bei 965 oder 21.5 Prozent der LG-Kandidaten bzw. bei 1’142 oder 31.7 Prozent der KG-Kandidaten der Fall gewesen.

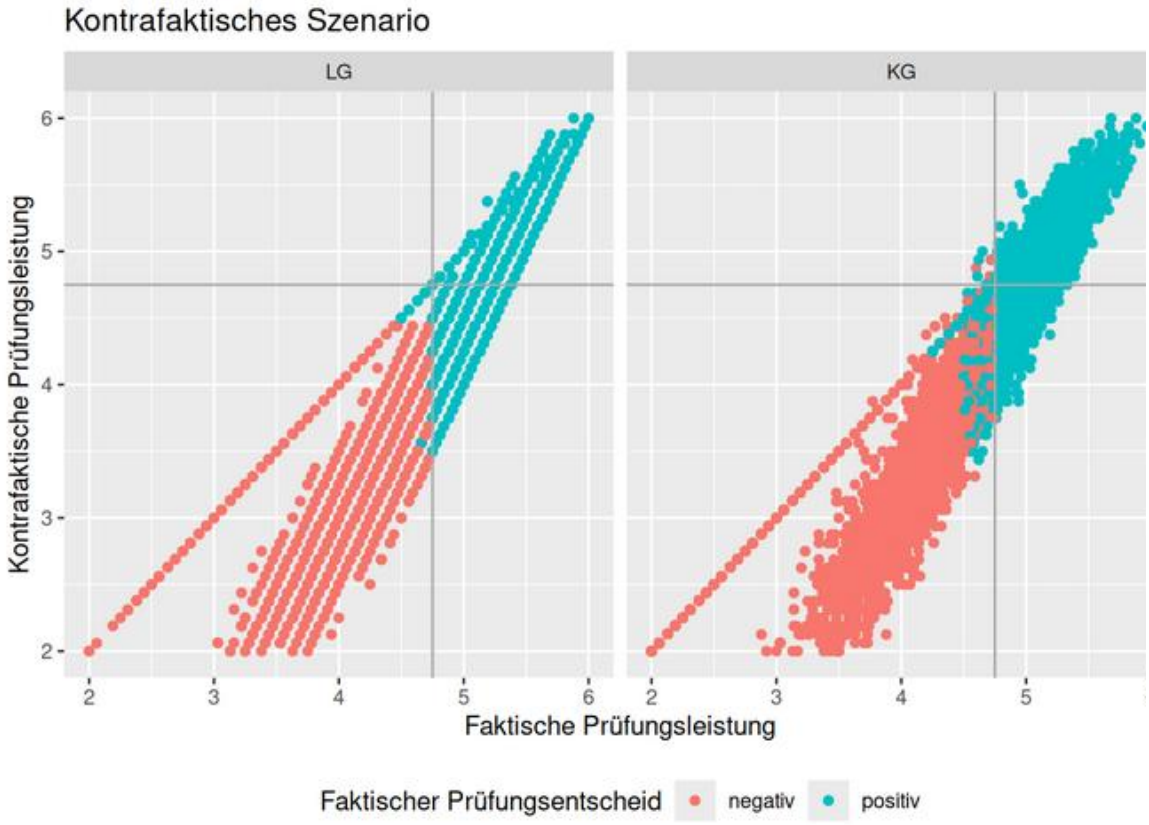
Im linken unteren Quadranten der Tabelle und auch im linken unteren Quadranten der Abbildung finden sich jene Jugendliche, die sowohl nach dem neuen Regime post-VAM (also faktisch) *nicht* bestanden haben und auch *nicht* bestanden hätten, wenn das alte Regime ante-VAM (also kontrafaktisch) gegolten hätte. Das wäre bei 2’089 oder 46.6

Prozent der LG-Kandidaten bzw. bei 1’984 oder 55.1 Prozent der KG-Kandidaten der Fall gewesen.

Rechts unten in der Tabelle und auch in der Abbildung finden sich jene Jugendliche, die nach dem neuen Regime post-VAM (also faktisch) bestanden haben, aber *nicht* bestanden hätten, wenn das alte Regime ante-VAM (also kontrafaktisch) gegolten hätte. Das wäre bei 1’433 oder 31.9 Prozent der LG-Kandidaten bzw. bei 464 oder 12.9 Prozent der KG-Kandidaten der Fall gewesen. Diese Jugendlichen haben so gesehen von dem neuen Regime profitiert, auch wenn diese Interpretation nur dann zulässig ist, wenn man von den eben dargestellten Annahmen ausgeht.

Abbildung 4

Kontrafaktische Prüfungstestung und faktische Prüfungsleistung des LG- und KG- Prüfungstyps nach faktischem Prüfungsentscheid



Links oben in der Tabelle und auch in der Abbildung finden sich schliesslich jene Jugendliche, die nach dem neuen Regime post-VAM (also faktisch) *nicht* bestanden haben, aber bestanden hätten, wenn das alte Regime ante-VAM (also kontrafaktisch) gegolten hätte. Für die LG-Prüfung würde das keine einzige Kandidatin bzw. keinen einzigen Kandidaten betreffen und für die KG-Prüfung lediglich zehn oder 0.3 Prozent der Kandidat*innen. Fast niemand hat also faktisch nicht bestanden, hätte aber bestanden, wenn die Vornoten nicht gezählt hätten. Dieser Vorteil für die Kandidat*innen scheint

zudem vor allem bei der LG- und der KG-Prüfung umso grösser zu sein, je schwächer die eigentliche Prüfungsleistung ist. Das ist daran zu erkennen, dass die faktische Prüfungsleistung im niedrigen Fähigkeitsbereich weiter von der gedachten Diagonalen in der Abbildung entfernt, ist als im oberen Fähigkeitsbereich.

Man kann das Szenario wie einen Vergleich mit zwei unterschiedlichen Regelbüchern verstehen, die auf dieselben beobachteten Prüfungsleistungen angewendet werden: Das «neue Regelbuch» kombiniert Prüfungsnote und Vornote zu einem Entscheid, dagegen würde das «alte Regelbuch» den Entscheid allein auf Basis der Prüfungsnote fällen. Im kontrafaktischen Teil wird also nicht behauptet, die Realität hätte sich tatsächlich so abgespielt; es ist ein rechnerisches Gedankenexperiment, das alles andere im neuen Prüfungsregime konstant hält (Prüfungsinhalt, Skalierung, Bestehensgrenzen, Anmeldeverhalten, Motivation usw.) und einzig die Entscheidungsregel austauscht. Genau deshalb sind die Ergebnisse als modellbasierte Einordnung zu lesen: Sie machen greifbar, wie stark die zusätzliche Komponente «Vornote» den Entscheid unter diesen Annahmen verschieben kann – ohne damit eine Aussage darüber zu treffen, wie sich die Realität unter einer anderen Regel tatsächlich entwickelt hätte.

Zusammengefasst zeigt sich, dass die Berücksichtigung der Vornote fast ausschliesslich mit einem günstigeren Entscheid für die Kandidat*innen einhergeht. Plausibel wird dieses Muster auch dadurch, dass Kandidat*innen sich im Durchschnitt mit hohen Vornoten zur Prüfung anmelden: So liegt die durchschnittliche Vornote in Mathematik bei der LG-Prüfung bei $M = 5.25$ selbst bei den Kandidat*innen, die am Ende nicht bestehen; bei den Bestehenden liegt sie mit $M = 5.67$ nochmals höher. Abschliessend sei noch einmal betont, dass hier ein sehr hypothetisches Szenario interpretiert wird. Die Resultate sind als transparente Illustration innerhalb der getroffenen Annahmen zu verstehen, nicht als endgültige Schlussfolgerung über «tatsächliche» Wirkungen der Vornote.

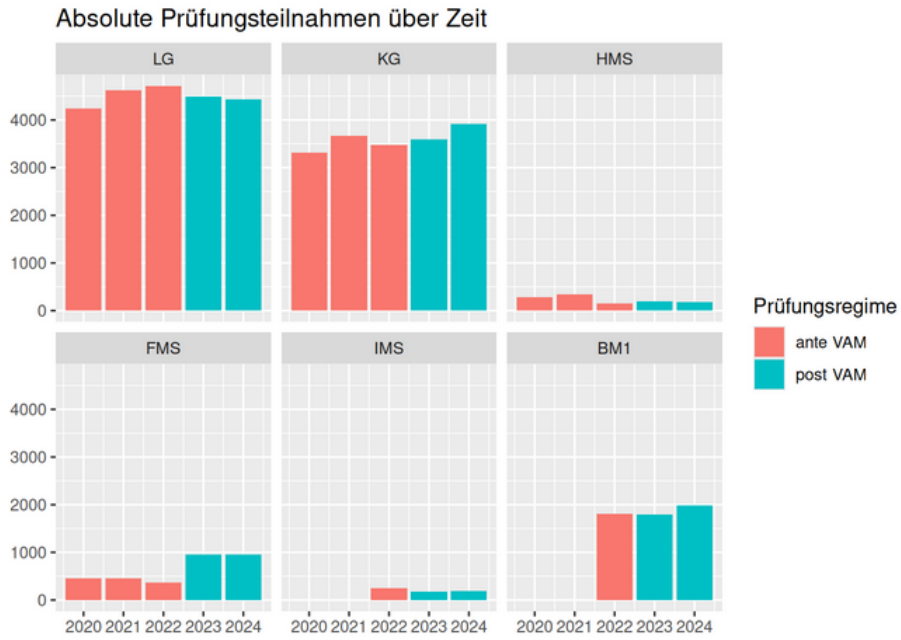
8. Anmeldezahlen, Bestehenszahlen und -quoten

Die Veränderungen der Anmeldezahlen, Bestehenszahlen und der daraus abgeleiteten Bestehensquoten über die Zeit sind in den folgenden Abbildungen dargestellt. Die Anmeldezahlen scheinen über die Zeit stabil zu bleiben. Lediglich für die FMS-Prüfung scheint es mehr als eine Verdopplung der Anmeldezahlen gegeben zu haben, so beispielsweise von 369 Fällen (2022) auf 951 Fälle (2023). Aufgrund der vorhandenen Datengrundlage muss jedoch zunächst unklar bleiben, ob es sich um ein einmaliges Phänomen handelt oder ob sich dieses Niveau in Zukunft verstetigen wird. Die absoluten Bestehenszahlen sind entweder stabil oder nehmen sogar zu, wie aus **Abbildung 5** ersichtlich. Vor allem bei der KG- und bei der FMS-Prüfung haben die Bestehenszahlen zum Teil deutlich zugenommen. Aus dem Quotienten der Anmelde- und der Bestehenszahlen ergibt sie die Bestehensquote, die in **Abbildung 6** zu sehen ist. Hier zeigen sich im Regimevergleich Zuwächse von einigen wenigen Prozentpunkten bei der LG-, der FMS- und der IMS-Prüfung und von knapp zehn Prozentpunkten bei der KG-

Prüfung. Bei der BM1-Prüfung sind die Bestehensquoten stabil geblieben und bei der HMS-Prüfung sind sie deutlich gesunken.

Abbildung 5

Prüfungsteilnahmen vor und nach dem Regimewechsel, aufgeteilt nach den einzelnen Prüfungstypen zwischen 2020 und 2024



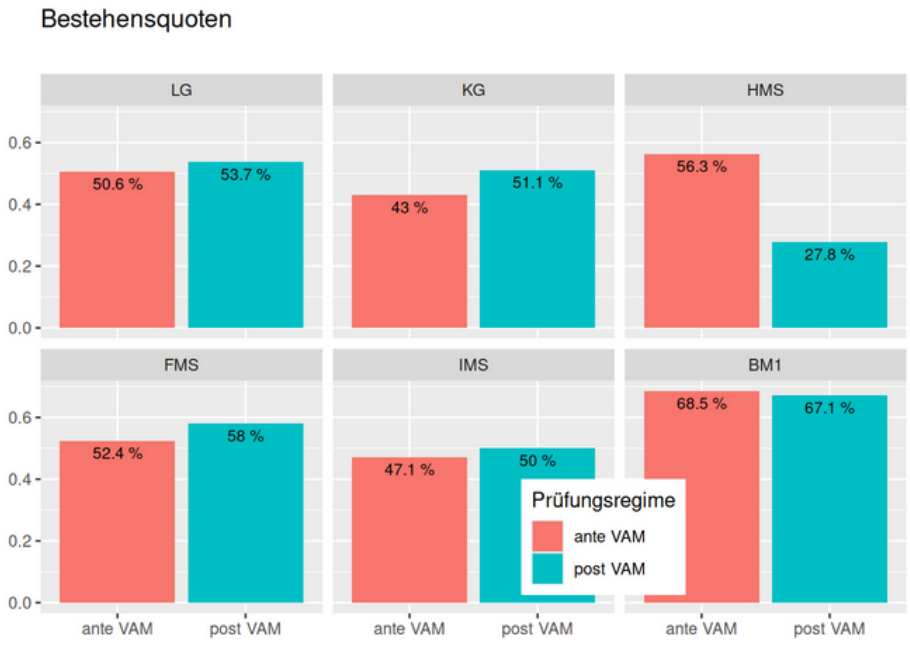
Bei dem zuletzt genannten Befund für die HMS-Prüfung müssen für dessen Interpretation jedoch drei Besonderheiten berücksichtigt werden, die diesen Einbruch in den Bestehensquoten erklären. Zunächst wurden für die bildungsstatistischen Auswertungen keine Personen, sondern eindeutig als bestanden oder nicht bestanden klassifizierte Prüfungsteilnahmen als Datengrundlage verwendet. Gleichzeitig war es post-VAM möglich, sich für die HMS als Ausweichoption im Rahmen der KG-Prüfung anzumelden. Wie in Abschnitt 10 ausführlicher berichtet, wurde diese Option auch von etwa der Hälfte der Kandidat*innen genutzt, letztlich aber nicht realisiert, weil das Prüfungsergebnis für die Anmeldung auf das KG ausreichend gewesen ist. In der zur Verfügung stehenden Bildungsstatistik wurden diese Fälle als «nicht bestanden» kodiert, weil es letztlich nicht zu einer Aufnahme auf die HMS gekommen ist. Somit sieht es lediglich so aus, als ob die Bestehensquoten dramatisch gesunken wären. Tatsächlich zeigt sich aber, wie weiter unten in **Abbildung 13** zu sehen ist, dass die Bestehensquoten bei den Doppelanmeldungen sogar angestiegen sind, wenn man die Kodierung des Bestehens entsprechend korrigiert.

Für die Interpretation aller hier vorgestellten Befunde ist zentral, dass die ausgewiesenen Anmelde- und Bestehenszahlen sowie die daraus abgeleiteten Bestehensquoten auf jener Teilmenge beruhen, für die ein eindeutiges Prüfungsergebnis vorliegt. Fälle ohne Ergebnis (etwa Nichterscheinen oder Abbruch) gehen nicht ein. Damit sind die hier berichteten Quoten «bedingt auf verwertbare Ergebnisse» zu lesen; Vergleiche zwischen

Jahren oder Prüfungstypen können zusätzlich beeinflusst werden, falls sich der Anteil solcher Fälle ohne Ergebnis verändert (Schafer & Graham, 2002).

Abbildung 6

Bestehensquoten vor und nach dem Regimewechsel, aufgeteilt nach den einzelnen Prüfungstypen



Zudem ist der Regimevergleich zeitlich asymmetrisch. Auffällige Veränderungen sollten deshalb zunächst als Befund für den beobachteten Jahrgang verstanden werden. Schliesslich gilt für Quoten grundsätzlich, dass sie bei kleineren Fallzahlen naturgemäss volatiler sind als bei grossen Prüfungstypen; für die inhaltliche Einordnung ist daher neben der Richtung auch die Unsicherheit der Quoten mitzudenken, etwa über Konfidenzintervalle für Anteilswerte (Brown et al., 2001). Diese Leitplanken ändern die Befunde nicht, helfen aber, ihre Tragweite und Vergleichbarkeit korrekt einzuordnen.

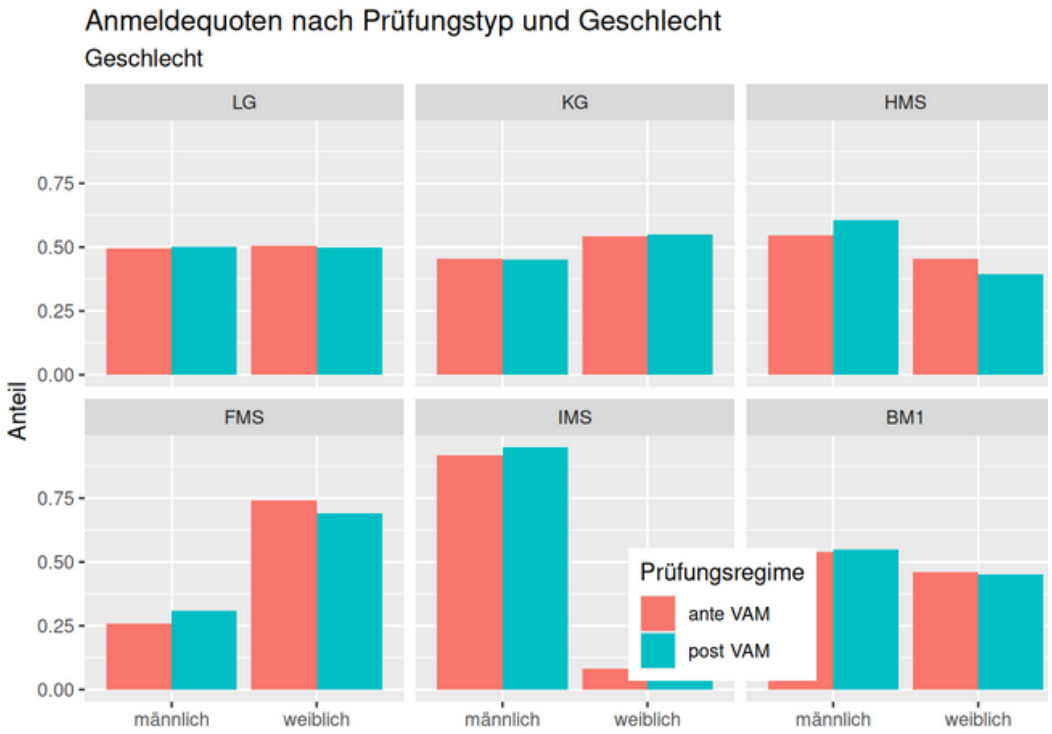
9. Anmeldezahlen und Bestehensquoten je nach Gruppenzugehörigkeit

Die im Folgenden untersuchte Frage bezog sich darauf, ob sich mit dem Regimewechsel die Anmeldequoten und die Bestehensquoten bei den einzelnen Prüfungstypen je nach Gruppenzugehörigkeit (z. B. Geschlecht oder Nationalität) geändert haben. Bei den Anmeldequoten gibt es keine wirklich auffälligen Befunde. Zwar ist beispielsweise der Anteil der deutschsprachigen Jugendlichen post-VAM (2023-2024) im Vergleich zu ante-VAM (2020-2022) leicht rückläufig. Beispielsweise sinkt der Anteil der Kandidierenden mit Muttersprache Deutsch in der LG-Prüfung von 79.7 auf 75.5 Prozent und in der KG-Prüfung von 75.9 auf 70.7 Prozent. Gleichzeitig nehmen die Anteile anderer

Sprachgruppen entsprechend leicht zu. Der Unterschied ist insgesamt aber nicht gross und lässt sich vermutlich grösstenteils über die veränderte Bevölkerungszusammensetzung erklären. Einzig beim Geschlecht scheint es einige Unterschiede zu geben, die es wert sind, genauer betrachtet zu werden. Wie in **Abbildung 7** zu erkennen ist, hat der Anteil der Knaben bei der HMS-Prüfung, der IMS-Prüfung und tendenziell auch bei der FMS-Prüfung leicht zugenommen.

Abbildung 7

Anmeldequoten vor und nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und dem Geschlecht



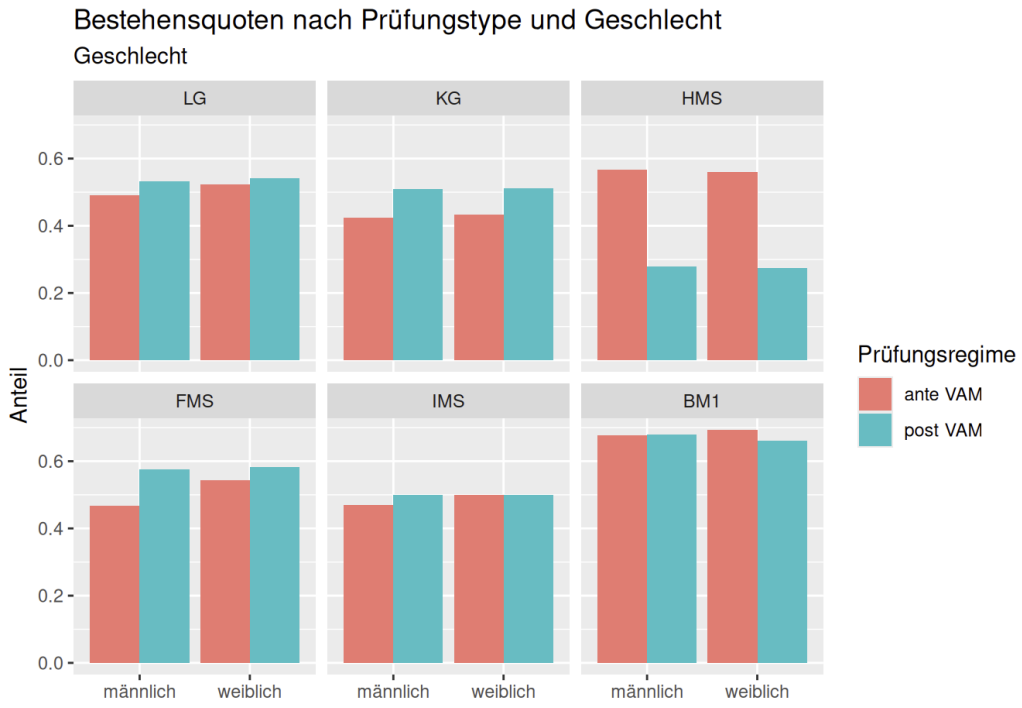
Bei den Bestehensquoten in Abhängigkeit von der Gruppenzugehörigkeit scheint es dagegen grössere Unterschiede zwischen ante-VAM (2020-2022) und post-VAM (2023-2024) zu geben. Die dazugehörigen Abbildungen sind auf den folgenden Seiten dargestellt, sofern es für bestimmte Gruppen Unterschiede zwischen den zwei Regimen gab. Hinsichtlich des Geschlechts fällt in **Abbildung 8** auf, dass die Bestehensquoten für Knaben bei der FMS-Prüfung überproportional zugenommen haben.

Bezogen auf die Nationalität ist das Bild etwas komplexer, wie aus **Abbildung 9** ersichtlich wird. Bei der KG-Prüfung scheinen sich Jugendliche mit Schweizer Nationalität besonders nach dem Regimewechsel stark verbessert zu haben, Jugendliche mit sonstiger europäischer Nationalität nicht so stark und Jugendliche mit aussereuropäischer Nationalität am wenigsten. Bei der HMS-Prüfung sind die Bestehensquoten für Jugendliche mit aussereuropäischer Nationalität sehr deutlich gesunken. Aber auch für Jugendliche mit Schweizer Nationalität haben sie sich etwa halbiert, während sie für

Jugendliche mit sonstiger europäischer Nationalität relativ gesehen nur wenig gesunken sind. Bei der FMS-Prüfung hat die Bestehensquote der Jugendlichen mit aussereuropäischer Nationalität am stärksten mit dem Regimewechsel zugenommen, während die Bestehensquoten für die zwei anderen Gruppen sehr viel weniger angestiegen sind.

Abbildung 8

Bestehensquoten vor und nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und dem Geschlecht



Eine Unterscheidung nach der Muttersprache gibt hier ein etwas differenziertes Bild, wie in **Abbildung 10** ersichtlich. Die eben beschriebenen Unterschiede nach Nationalität bei der KG-Prüfung scheint nämlich *nicht* durch den sprachlichen Hintergrund der Jugendlichen erklärbar zu sein, weil hier die Veränderungen des Regimewechsels alle vier Sprachgruppen in etwa gleich betreffen. Bei der HMS-Prüfung haben sich die Bestehensquoten für Jugendliche mit deutscher oder einer nichteuropäischen Muttersprache in etwa halbiert, während der Rückgang für Jugendliche mit sonstiger germanischer oder romanischer Muttersprache und für Jugendliche mit sonstiger europäischer Muttersprache nicht so stark ausgeprägt war. Bei der FMS-Prüfung ist die Bestehensquote der Jugendlichen mit einer nichteuropäischen Muttersprache am stärksten mit dem Regimewechsel gestiegen, was mit dem oben beschriebenen Befund zur aussereuropäischen Nationalität korrespondiert. Und bezüglich der IMS-Prüfung zeigt sich etwas differenzierter als bei der Auswertung nach Nationalität, dass vor allem bei Jugendlichen mit einer nichtgermanischen und nichtromanischen Muttersprache der stärkste Anstieg der Bestehensquoten zu verzeichnen war.

Abbildung 9

Bestehensquoten vor und nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und der Nationalität

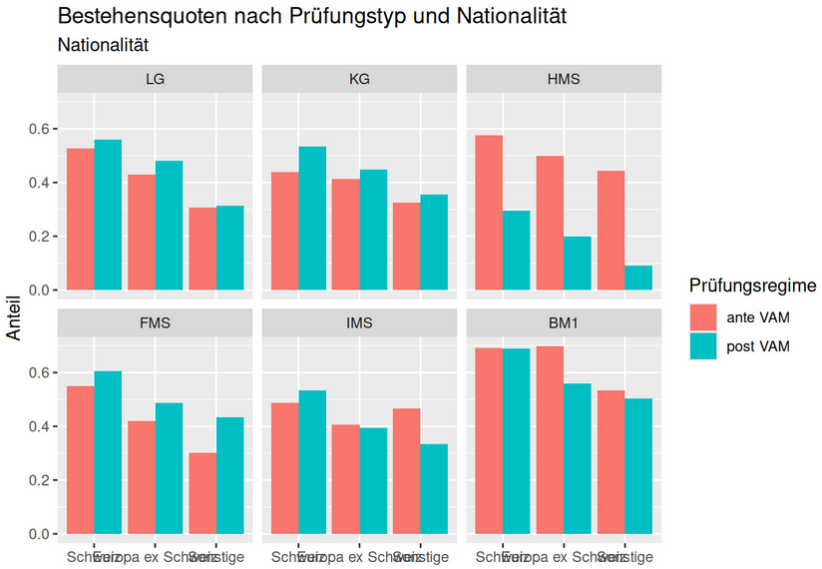
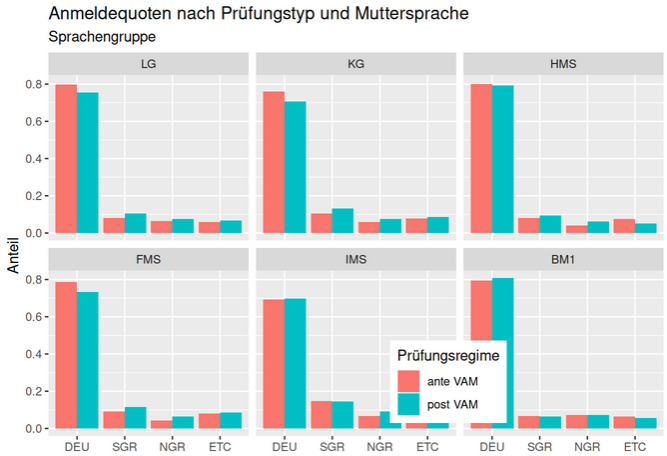


Abbildung 10

Bestehensquoten vor und nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und der Sprachgruppe

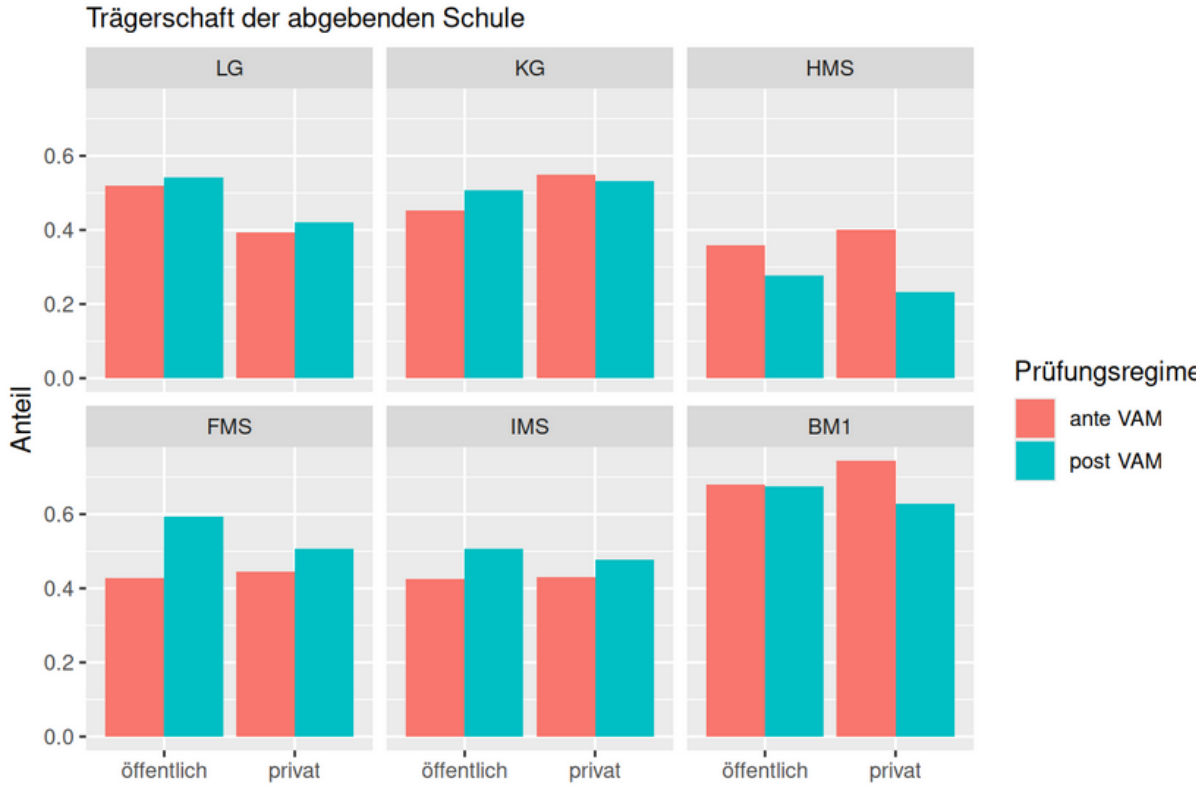


In den vorliegenden Auswertungen auf **Abbildung 11** zeigen sich nach Trägerschaft für einzelne Prüfungstypen Verschiebungen in den Bestehensquoten, während für die Veränderungen der Prüfungsnoten nach Trägerschaft keine auffälligen differenziellen Muster berichtet wurden. Diese Gegenüberstellung kann zunächst als Unterschied zwischen einer kontinuierlichen Messgrösse (Prüfungsnote) und einer dichotomen Klassifikationsentscheidung (Bestehen/Nichtbestehen) verstanden werden: Quoten reagieren besonders sensibel, wenn ein relevanter Anteil der Kandidat*innen nahe an einer Bestehensgrenze liegt oder sich die Lage dieser Grenze im Verhältnis zur Verteilung verändert, ohne dass sich dies zwingend in gruppenspezifisch unterschiedlichen

Mittelwertveränderungen der Prüfungsnote niederschlagen muss. Zudem geht die Abbildung einer kontinuierlichen Skala auf eine binäre Entscheidung grundsätzlich mit Informationsverlust einher; dadurch können Muster bei Bestehensquoten anders sichtbar werden als bei Analysen auf der kontinuierlichen Skala, selbst wenn die zugrunde liegenden Notenveränderungen ähnlich ausfallen (Altman & Royston, 2006; MacCallum et al., 2002). Diese Hinweise dienen der Einordnung des Befunds und der konsistenten gemeinsamen Interpretation von Noten- und Bestehensanalysen; sie sind nicht als Ursachenanalyse zu verstehen.

Abbildung 11

Bestehensquoten vor und nach dem Regimewechsel, getrennt nach den einzelnen Prüfungstypen und Trägerschaft (nur 2022-2024)



Insgesamt zeigt sich im Regimevergleich ein konsistentes Grundmuster: Grössere Veränderungen sind eher bei den Bestehensquoten als bei den Anmeldequoten zu beobachten. Gleichzeitig sind die berichteten Unterschiede nicht als einheitliches, über alle Prüfungstypen hinweg gleichgerichtetes Gruppenmuster zu verstehen. Vielmehr treten differenzielle Veränderungen punktuell auf – je nach Prüfungstyp und je nach betrachteter Gruppenvariable (Geschlecht, Nationalität, Muttersprache, Trägerschaft) – und sie betreffen die Veränderung der Quoten zwischen ante-VAM (2020-2022) und post-VAM (2023-2024), nicht zwingend das generelle Niveau der Gruppen. Diese Lesart hilft auch, scheinbar widersprüchliche Befunde nebeneinander einzuordnen: Eine Gruppe

kann sich in einem Prüfungstyp deutlich «verbessern», ohne dass sich daraus eine generelle Aussage über alle Prüfungstypen oder über die Anmeldequoten ableiten liesse.

Für die Interpretation sind mehrere methodische Leitplanken mitzudenken. Erstens werden hier viele Kombinationen ausgewertet (mehrere Prüfungstypen × mehrere Gruppenvariablen), wodurch einzelne Unterschiede auch zufällig auftreten können; besonders belastbar sind Befunde, die innerhalb eines Prüfungstyps konsistent sind oder sich in verwandten Gruppierungen in ähnlicher Richtung zeigen (Benjamini & Hochberg, 1995). Zweitens können bei Prüfungstypen mit kleinen Anmeldezahlen prozentuale Veränderungen gross wirken, obwohl die absoluten Fallzahlen gering sind; die Grössenordnung sollte daher immer im Licht der zugrunde liegenden Stichprobengrösse und der Unsicherheit von Anteilswerten gelesen werden (Brown et al., 2001). Drittens ist der Regimevergleich zeitlich asymmetrisch (post-VAM derzeit nur zwei Jahrgänge), sodass allfällige Veränderungen zunächst als Befund für den beobachteten Jahrgang zu verstehen sind. Viertens ist bei einzelnen Gruppenvariablen – insbesondere der Trägerschaft – zu berücksichtigen, dass fehlende Angaben die Datengrundlage einschränken können; die Befunde sind dann als Hinweise innerhalb der verfügbaren Teilmenge zu lesen (Schafer & Graham, 2002). Zusammengenommen stützen die Analysen damit eine datennahe Schlussfolgerung: Die Veränderungen der Bestehensquoten im Regimevergleich sind real und teils gruppenspezifisch sichtbar, sollten aber konsequent als prüfungstyp- und datengrundlagenabhängige Muster interpretiert werden – nicht als generelle Verschiebung «für» oder «gegen» bestimmte Gruppen.

10. Doppelanmeldungen KG/HMS

Bei der Anmeldung zur KG-Prüfung konnten die Jugendlichen bzw. deren Eltern auch angeben, ob sie sich im Falle des Nichtbestehens der KG-Prüfung aber bei genügenden Prüfungsleistungen für die HMS-Option auch für letztere anmelden möchten. Der Anteil der Jugendlichen, die sich für die HMS-Option entschieden haben, ist in **Tabelle 2** dargestellt. Es wird deutlich, dass die HMS-Option bis 2022 von knapp unter 30 Prozent der Jugendlichen ausgewählt worden ist und dann 2023 mit dem Regimewechsel auf etwa 50 Prozent angestiegen ist. Konkret liegt der Anteil der Kandidaten mit HMS-Option in den Jahren 2020-2022 jeweils bei rund drei Zehnteln und steigt 2023 auf gut die Hälfte (2020: 29.3 Prozent; 2021: 27.8 Prozent; 2022: 27.8 Prozent; 2023: 48.3 Prozent; 2024: 53.4 Prozent). Ob es sich bei diesem Anstieg um einmalige jahrgangsspezifische Ausreisserwerte handelt oder ob sich der Trend in Zukunft fortsetzen wird, müssen zukünftige Untersuchungen zeigen.

Unterschiede betrugen nur wenige Prozentpunkte und waren bei den Gruppenvariablen «Geschlecht» und «Trägerschaft» am höchsten ausgeprägt. So nutzten männliche Jugendliche und solche, die von einer Schule in privater Trägerschaft kamen, die HMS-Option etwas häufiger als weibliche Jugendliche und solche, die von einer Schule in

öffentlicher Trägerschaft kamen. Diese Zahlen beziehen sich auf alle vier untersuchten Jahrgänge.

Tabelle 2

Anteil der HMS- Option nach Prüfungsjahr

Verteilung der HMS-Option über die Zeit

Prüfungsjahr	HMS-Option	Fälle	Anteil im Prüfungsjahr
2020	ja	969	0.29
2020	nein	2341	0.71
2021	ja	1019	0.28
2021	nein	2649	0.72
2022	ja	969	0.28
2022	nein	2512	0.72
2023	ja	1739	0.48
2023	nein	1861	0.52
2024	ja	2092	0.53
2024	nein	1828	0.47

Tabelle 3

Wahl der HMS- Option nach Geschlecht, Muttersprache, Trägerschaft der abgebenden Schule, Nationalität und Aufenthaltsdauer

Wahl der HMS-Option nach Geschlecht

Geschlecht	HMS ja	HMS nein
männlich	41%	59%
weiblich	35%	65%

Wahl der HMS-Option nach Muttersprache

Muttersprache	HMS ja	HMS nein
DEU	38%	62%
SGR	39%	61%
NGR	36%	64%
ETC	40%	60%

Wahl der HMS-Option nach Aufenthaltsdauer

Aufenthaltsdauer	HMS ja	HMS r
bis zu 10 Jahre	44%	56%
mehr als 10 Jahre	45%	55%
einheimisch	36%	64%

Wahl der HMS-Option nach Nationalität

Nationalität	HMS ja	HMS nein
Schweiz	38%	62%
Europa ex Schweiz	34%	66%
Sonstige	38%	62%

Wahl der HMS-Option nach Trägerschaft der abgebenden Schule

Trägerschaft	HMS ja	HMS nein
öffentlich	43%	57%
privat	49%	51%

Abbildung 12

Anmeldequoten nach HMS- Option vor und nach dem Regimewechsel

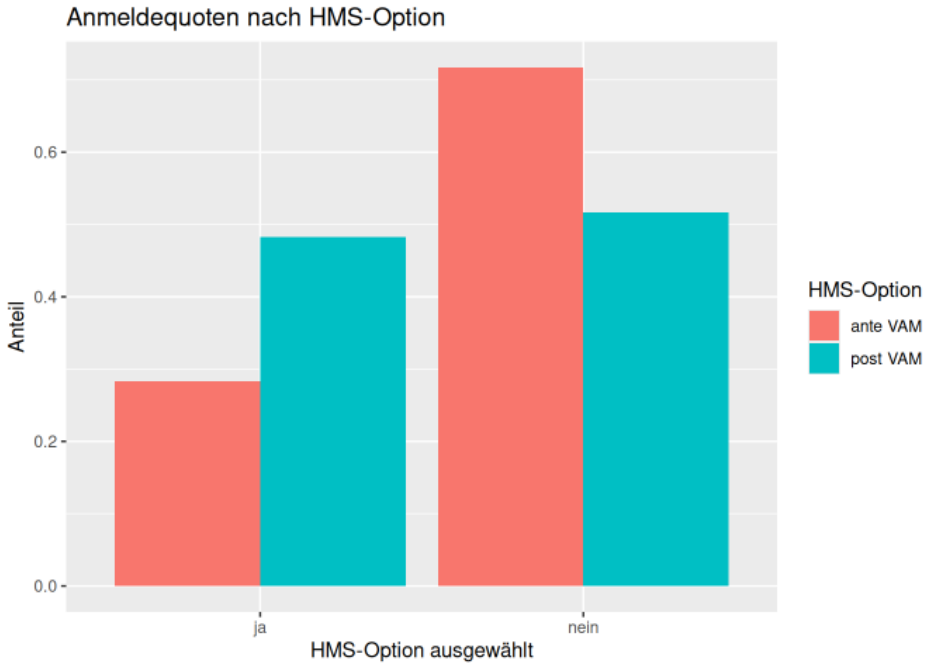
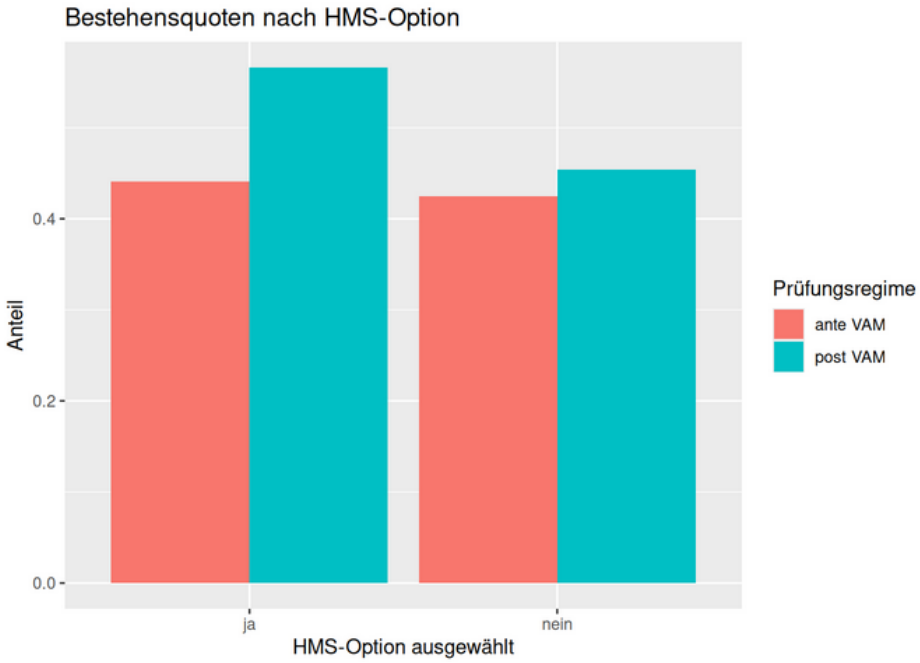


Abbildung 13

Bestehensquoten nach HMS-Option vor und nach dem Regimewechsel



In einem nächsten Schritt wurde untersucht, ob die Gruppenzugehörigkeit (z. B. Geschlecht oder Nationalität) mit der Wahl der HMS-Option zusammenhängt. Hier gab es keine auffälligen Befunde, wie aus **Tabelle 3** zu entnehmen ist. Die grössten

Schliesslich wurde verglichen, wie sich mit dem Regimewechsel die Anmelde- und Bestehensquoten für diejenigen KG-Prüfungskandidaten mit und ohne HMS-Option verändert haben. In einer ersten Auswertung wurde zunächst betrachtet, ob der Regimewechsel eine Auswirkung auf die Auswahl der HMS-Option hatte. **Abbildung 12** zeigt, dass die Quote derjenigen, die bei der KG-Prüfung auch die HMS-Option ausgewählt haben, von 28.3 auf 50.9 Prozent angestiegen ist und sich damit fast verdoppelt hat.

Auch zeigt sich, dass es bei den Bestehensquoten mit dem Regimewechsel für diejenigen Jugendlichen zu einer Erhöhung der Bestehenswahrscheinlichkeit gekommen ist, wenn sie die HMS-Option gewählt haben. Dieser Zusammenhang ist in **Abbildung 13** dargestellt. Insgesamt ist die Bestehensquote bei der KG-Prüfung ja sowieso angestiegen. Der Anstieg war aber grösser für diejenigen, die die HMS-Option ausgewählt haben. Die naheliegendste Interpretation für diesen Befund ist, dass die HMS-Option von fähigeren Kandidat*innen ausgewählt worden ist. Es könnte aber auch sein, dass Jugendliche, die die HMS-Option gewählt haben, anders an die KG-Prüfung herangegangen sind (etwa selbstsicherer, da sie wussten, dass es zwei Möglichkeiten für einen Maturitätsabschluss gibt) und sie deswegen bessere Leistungen erzielt haben.

Die hier ausgewiesenen Doppelanmeldungen sind vor allem als Hinweis darauf zu verstehen, dass in den administrativen Kennwerten zwei Bezugsgrössen nebeneinander existieren können: eine anwendungsbezogene Perspektive (jede Anmeldung zählt) und eine personenbezogene Perspektive (jede*r Kandidat*in zählt). Doppelanmeldungen führen dazu, dass Anmeldezahlen und – je nach Auswertungslogik – auch Bestehenszahlen in einzelnen Prüfungstypen nicht zwingend der Zahl einzigartiger Kandidat*innen entsprechen. Für die Interpretation der vorangegangenen Quoten ist daher wichtig, ob sie als Quoten pro Anmeldung (prüfungstypbezogen) oder als Quoten pro Kandidat*in (personenbezogen über Prüfungstypen hinweg) zu lesen sind. In einem System mit Doppelanmeldungen können beide Sichtweisen sinnvoll sein, sie beantworten aber unterschiedliche Fragen: Die anwendungsbezogene Sicht beschreibt die Erfolgswahrscheinlichkeit innerhalb eines Prüfungstyps, die personenbezogene Sicht beschreibt die Erfolgswahrscheinlichkeit einer Kandidatin bzw. eines Kandidaten unter Berücksichtigung möglicher Mehrfachanmeldungen.

Zusätzlich ist zu beachten, dass die Identifikation von Doppelanmeldungen eine Wiedererkennung derselben Person über Anmeldungen hinweg voraussetzt. Wo das nicht gegeben ist, können Doppelanmeldungen unter- oder überschätzt werden. Entsprechend sind die Befunde zu Doppelanmeldungen primär als Transparenzinformation zur Datenstruktur zu lesen: Sie helfen, die Reichweite von anmeldebezogenen Kennwerten korrekt einzuordnen und zu entscheiden, wann ergänzende personenbezogene Auswertungen sinnvoll wären (Harron et al., 2017).

11. Zusammenfassende Betrachtung

Die Ergebnisse der bildungsstatistischen Auswertungen lassen sich wie folgt zusammenfassen, wenn man sich auf die wesentlichen Ergebnisse fokussiert. Vor dem Hintergrund der sowieso gestiegenen Prüfungsdurchschnitte post-VAM (2023-2024) im Vergleich zu ante-VAM (2020-2022) hat sich gezeigt, dass die Einführung des neuen Prüfungsregimes zu keiner systematischen Bevorzugung oder Benachteiligung einzelner Gruppen geführt hat. Es gibt zwar für einzelne Prüfungstypen und Prüfungsfächer einzelne Gruppen, die ante-VAM und post-VAM unterschiedliche Muster in den Mittelwerten hatten. Aber dieser Fall ist selten und vor allen Dingen muss mit den vorhandenen Daten unklar bleiben, ob es sich dabei lediglich um jahrgangsspezifische Phänomene handelt, die sich auf mittlere Sicht wieder verlieren könnten.

Eines der wesentlichen Ergebnisse dieser Evaluation war, dass die Vornoten und die Prüfungsnoten nur mässig miteinander korreliert sind. Mit anderen Worten werden hier unterschiedliche Fähigkeiten gemessen. Es ist plausibel anzunehmen, dass sich in den Vornoten vor allem allgemeine kognitive Fähigkeiten sowie die Fähigkeit zur dauerhaften Leistung niederschlägt, während in den Prüfungsnoten die Fähigkeit abgebildet wird, kurzfristig eine hohe Leistung zu zeigen, wenn es darauf ankommt (vgl. Napolitano et al., 2021). Es wird sich in zukünftig geplanten Auswertungen zeigen, ob Vornoten oder Prüfungsnoten besser dazu geeignet sind, den Erfolg in einer Maturitätsschule, wie er sich beispielsweise im erfolgreichen Bestehen der Probezeit widerspiegelt, besser vorherzusagen (vgl. dazu Kapitel 2, Abschnitt 7).

Bezüglich der Streuung der Vornoten über die verschiedenen Schulen und regionalen Einheiten hinweg hat sich gezeigt, dass es einen kleinen Unterschied macht, welches die abgebende Schule ist, nicht aber in welchem Schulbezirk diese liegt. Zwar unterscheiden sich die Mittelwerte der Vornoten zum Teil beträchtlich zwischen den einzelnen Schulen, aber die Streuung innerhalb der Schulen ist wesentlich grösser als die Streuung zwischen den Schulen. Zudem unterscheiden sich die Intraclasskorrelationen zwischen den einzelnen Schulfächern. Alle diese Befunde sprechen dafür, dass die Vornoten in allererster Linie individuelle Leistungen widerspiegeln und so gut wie gar nicht die Zugehörigkeit zu einer bestimmten abgebenden Schule. Im Sinne der Chancengleichheit scheint zumindest in dieser Hinsicht also nichts gegen die Verwendung von Vornoten als Indikator für die individuelle Leistungsfähigkeit zu sprechen.

Wie man erwarten würde, unterscheiden sich die Vornoten von Jugendlichen, die bestanden haben, von denen, die abgelehnt wurden, im Durchschnitt zum Teil deutlich. Gleichzeitig findet sich eine hohe Überlappung beider Verteilungen, was dafürspricht, dass sich in dem gegenwärtigen Regime auch schwächere Vornoten durch eine gute Prüfungsleistung ausgleichen lassen. Wichtig ist an dieser Stelle zu betonen, dass damit noch keine Aussage darüber gemacht wird, wie gut sich dieser Indikator dazu eignet, den Schulerfolg auf der weiterführenden Schule vorherzusagen.

Bei der Betrachtung eines hypothetischen kontrafaktischen Szenarios, bei dem die Vornoten nicht für das Bestehen der Prüfung zählen würden, hat sich auf den ersten Blick gezeigt, dass durch die Berücksichtigung der Vornoten eher mehr Kandidat*innen zugelassen worden sind, als das ohne die Berücksichtigung der Vornoten der Fall gewesen wäre. Diese Interpretation ist jedoch nur dann zulässig, wenn man davon ausgehen könnte, dass sich insbesondere bezüglich der Skalierung der Prüfungen nichts zwischen ante-VAM (2020-2022) und post-VAM (2023-2024) bzw. auch zwischen den einzelnen Prüfungsjahrgängen nichts geändert hätte. Das ist selbstverständlich eine Annahme, die in der Realität wahrscheinlich nicht haltbar ist, sodass dieser Befund eher als ein Ausgangspunkt für weitere Auswertungen, denn als endgültig interpretierbares Ergebnis gelesen werden sollte.

Im Grossen und Ganzen zeigt sich, dass die Anmelde- und Bestehenszahlen und die daraus abgeleiteten Bestehensquoten grösstenteils stabil geblieben sind. Gewisse Verschiebungen gibt es allerdings, wenn man die Bestehensquoten in Abhängigkeit zur Gruppenzugehörigkeit untersucht. Auch wenn es bezüglich der Prüfungsnoten keine systematische Bevorzugung oder Benachteiligung einer einzelnen Gruppe zu geben scheint, sollte beobachtet werden, ob sich in Zukunft deutlichere Verschiebungen zwischen den unterschiedlichen Gruppen entwickeln.

Schliesslich wurden die Doppelanmeldungen KG/HMS untersucht, wobei sich hier herausgestellt hat, dass die Zahl der Doppelanmeldung mit dem Regimewechsel deutlich zugenommen hat. Dabei scheint es so zu sein, dass diese Option von fähigeren Kandidat*innen ausgewählt worden ist. Ob es sich dabei um eine nachhaltige Entwicklung handelt, müssen Daten zukünftiger Prüfungsjahrgänge zeigen.

Methodisch sind die berichteten Ergebnisse durchgängig als deskriptive Befunde im Vergleich ante-VAM (2020–2022) versus post-VAM (2023-2024) zu lesen. Zwei Aspekte sind dabei wiederkehrend zentral. Erstens ist der Vergleich zeitlich asymmetrisch, da post-VAM (2023-2024) derzeit nur zwei Jahrgänge umfasst; solche Veränderungen können deshalb jahrgangsspezifisch sein und sollten erst mit weiteren Jahrgängen als stabil oder nachhaltig beurteilt werden. Zweitens sind Mittelwerte und Effektstärken über Prüfungsjahre hinweg zwar auf derselben Notenskala ausgewiesen, dennoch handelt es sich nicht um identische Testformen; ohne explizite Verknüpfung der Testformen (Equating/Linking) sind Vergleiche über Prüfungseditionen hinweg als Vergleiche beobachteter Jahrgänge zu interpretieren, nicht als streng formgleiche Veränderungen eines konstanten Messinstruments (Dorans et al., 2010).

Hinzu kommen drei wiederkehrende Punkte zur Robustheit von Unterschieden. Erstens werden in mehreren Abschnitten viele Gruppen- und Prüfungstypkombinationen betrachtet; einzelne auffällige Unterschiede sind deshalb erwartbar, ohne dass dahinter zwingend ein systematisches Muster stehen muss. Entsprechend sollte Befunden besonderes Gewicht gegeben werden, wenn sie innerhalb eines Prüfungstyps konsistent sind oder sich inhaltlich in verwandten Gruppierungen spiegeln (Benjamini & Hochberg, 1995). Zweitens sind Anteilswerte (Bestehensquoten) bei kleinen Fallzahlen naturgemäss

volatiler; für die inhaltliche Einordnung ist daher neben der Richtung auch die Unsicherheit der Quoten mitzudenken (Brown et al., 2001). Drittens beziehen sich mehrere Kennwerte auf Teilmengen mit gültigen Angaben beziehungsweise eindeutigem Ergebnis; fehlende Werte und Fälle ohne eindeutiges Ergebnis können die Vergleichbarkeit zwischen Jahren, Prüfungstypen und Gruppen beeinflussen und sollten deshalb bei der Interpretation als Rahmenbedingung konsequent mitgeführt werden (Schafer & Graham, 2002).

Literaturverzeichnis

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. BMJ, 332, 1080. doi:10.1136/bmj.332.7549.1080

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. Statistical Science, 16, 101–133. doi:10.1214/ss/1009213286

Bosker, R. J., & Witziers, B. (1996). A meta analytical approach regarding school effectiveness: The true size of school effects and the effect size of educational leadership (ERIC Document Reproduction Service No. ED392147). Department of Education at the University of Twente.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. ETS Research Report Series, 2010, i–41. doi:10.1002/j.2333-8504.2010.tb02236.x

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861–874. doi:10.1016/j.patrec.2005.10.010

Gaab, J. (2009). PASA – Primary Appraisal Secondary Appraisal: Ein Fragebogen zur Erfassung von situationsbezogenen kognitiven Bewertungen. Verhaltenstherapie, 19, 114–115. doi:10.1159/000223610

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology, 46, 1699–1710. doi:10.1093/ije/dyx177

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29, 60–87. doi:10.3102/0162373707299706

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7, 19–40. doi:10.1037/1082-989X.7.1.19

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. CBE—Life Sciences Education, 12, 345–351. doi:10.1187/cbe.13-04-0082

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30–46. doi:10.1037/1082-989X.1.1.30

Napolitano, C. M., Sewell, M. N., Yoon, H. J., Soto, C. J., & Roberts, B. W. (2021). Social, emotional, and behavioral skills: An integrative model of the skills associated with success during adolescence and across the life span. Frontiers in Education, 6, 679561. doi:10.3389/feduc.2021.679561

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147–177. doi:10.1037/1082-989X.7.2.147

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. Anesthesia and Analgesia, 126, 1763–1768. doi:10.1213/ANE.0000000000002864

Stockford, S. M. (2009). Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement [Doctoral dissertation]. Arizona State University.

Tomasik, M. J., Silbereisen, R. K., & Pinquart, M. (2010). Individuals negotiating demands of social change: A control-theoretical approach. European Psychologist, 15, 246–259. doi:10.1027/1016-9040/a000064

Kapitel 2: Onlinebefragung der Kinder und Jugendlichen

Martin J. Tomasik

Inhaltsverzeichnis

1. Einleitung

Die in Kapitel 1 dargestellten Befunde basieren auf administrativen Daten und beschreiben Unterschiede im Ergebnis der Prüfung. [...] Damit bleibt jedoch offen, wie die Prüfungssituation von den Jugendlichen selbst erlebt wird. In den folgenden Abschnitten des Kapitel 2 werden deshalb die Befunde der Onlinebefragung dargestellt. Diese Darstellung erfolgt entlang folgender Leitfragen:

- Wie erleben die Jugendlichen die ZAP? (vgl. Abschnitt 3)
- Haben die Jugendlichen einen öffentlichen und/oder privaten Vorbereitungskurs besucht? (vgl. Abschnitt 4)
- Erhielten sie ausserhalb von Schule und Vorbereitungskursen Unterstützung bei der Vorbereitung und wenn ja, von wem? (vgl. Abschnitt 5)
- War aus der Sicht der Jugendlichen der Prüfungsstoff mit dem Unterricht an der Volksschule abgedeckt? (vgl. Abschnitt 6)

Darüber hinaus werden in diesem Abschnitt Ergebnisse zum Bestehen der Probezeit vorgestellt (vgl. Abschnitt 7). Es wurde untersucht, ob sich das Bestehen der Probezeit durch die Vor- und Prüfungsnoten, die jeweilige Regionaleinheit (Schule, Schulgemeinde und Bezirk) und den Besuch von öffentlichen bzw. privaten Vorbereitungskursen vorhersagen lässt. Ausserdem wird der Zusammenhang mit den Prüfungsnoten vor und nach dem Regimewechsel dargestellt. Anschliessend erfolgt eine zusammenfassende Betrachtung aller Ergebnisse (vgl. Abschnitt 8).

Mit der Onlineumfrage soll einerseits die Perspektive der Jugendlichen näher untersucht werden. Andererseits sollen diese Angaben im Schlussbericht mit dem Prüfungsergebnis bzw. mit dem Bestehen der Probezeit in Verbindung gebracht werden, um Zusammenhänge zwischen den dort erhobenen Variablen und einem erfolgreichen Übertritt auf eine Maturitätsschule zu identifizieren. In diesem Bericht werden die wesentlichen Befunde zusammengefasst. [...]

Bei der Interpretation der Ergebnisse ist zu beachten, dass es sich bei der Datengrundlage um korrelative Daten handelt, aus der sich in der Regel keine kausalen Schlüsse ziehen lassen. Auch wenn es manchmal naheliegend wäre, von einem Ursache-Wirkungs-Verhältnis und nicht bloss von einem Zusammenhang auszugehen, der beispielsweise auch durch Drittvariablen verursacht worden sein könnte, ist diese Interpretation streng genommen nicht zulässig.

2. Methode, Stichprobe und Skaleneigenschaften

Die Daten für diese Auswertungen wurden mittels eines Onlinefragebogens auf der Plattform «Tivian» erhoben. Alle Jugendlichen, die sich im Frühjahr 2024 zu einer Aufnahmeprüfung (LG; KG; HMS und FMS, im Folgenden zusammengefasst zu MS für Mittelschulen; BM) angemeldet haben, wurden zu der Umfrage eingeladen. Die

Population bestand aus 11’297 Jugendlichen, davon haben sich 4’333 für die LG-Prüfung, 3’876 für die KG-Prüfung, 134 für die HMS-Prüfung, 982 für die FMS-Prüfung und 1’972 für die BM-Prüfung angemeldet.

Jugendliche, die sich für die LG-, KG- oder die HMS-Prüfung angemeldet haben, wurden am 05. März 2024 eingeladen, und Jugendliche, die sich für die FMS- oder BM-Prüfung angemeldet haben, am 07. März 2024. Die Einladung erfolgte per E-Mail an die Adresse, die die Jugendlichen bzw. deren Eltern auch für die Anmeldung zur Prüfung verwendet hatten. Alle Jugendlichen, die den Fragebogen noch nicht abgeschlossen haben, wurden am 12. März 2024 noch einmal erinnert. Die Umfrage wurde jeweils um Mitternacht am 18. März 2024 (für die LG-, KG- und HMS-Prüfung) bzw. am 20. März 2024 (für die FMS- und BM-Prüfung) geschlossen. Die Wahl des Befragungszeitraums stellte sicher, dass alle Kinder und Jugendlichen die Prüfungssituation selbst erlebt haben, aber niemandem das Prüfungsergebnis bekannt sein konnte. Somit wurde sichergestellt, dass die Einschätzungen und Selbstberichte nicht durch die Kenntnis des Prüfungsergebnisses beeinflusst worden sein konnten.

Tabelle 4

Anteil der abgeschlossenen Umfragen nach den einzelnen Prüfungstypen

Verteilung der Prüfungstypen

Prüfung	N
LG	1781
KG	1184
BM	407
MS	229
Total	3601

Die durchschnittliche Bearbeitungsdauer des Onlinefragebogens betrug *M* = 17.86 (*SD* = 8.35) Minuten, wobei es hier zwischen den Prüfungstypen Unterschiede gab (LG: *M* = 19.60, *SD* = 8.06; KG: *M* = 16.67, *SD* = 8.48; MS: *M* = 14.32, *SD* = 7.32; BM: *M* = 15.64, *SD* = 8.00). Die Bearbeitungszeit erklärt sich auch dadurch, dass alle nichtoffenen Fragen Pflichtfelder waren, die ausgefüllt werden mussten, um die Umfrage abzuschliessen. Erst dann konnten die Teilnehmenden an einer Verlosung teilnehmen, bei denen 500 mal 20 Franken verlost wurden.

An der Onlineumfrage haben am Ende insgesamt *N* = 3’601 Jugendliche teilgenommen, davon nach eigenen Angaben 1’649 Jungen, 1’939 Mädchen sowie 13 Sonstige (LG: *n* = 1’781; KG: *n* = 1’184; BM: *n* = 407; HMS/FMS/IMS: *n* = 229).

Die Verteilung der abgeschlossenen Umfragen auf die einzelnen Prüfungstypen ist in **Tabelle 4** dargestellt. Demnach liegen die Beendigungsquoten für das LG bei 41.1 Prozent, für das KG bei 30.6 Prozent, für die BM bei 20.6 Prozent und für die MS bei 20.5

Prozent. Die höheren Beendigungsquoten für jüngere Jugendliche erklären sich vielleicht damit, dass für sie die verlosten 20 Franken einen höheren subjektiven Wert hatten.

Von den insgesamt N = 3'601 Jugendlichen, die an der Onlineumfrage teilgenommen haben, konnten 3'496 Jugendliche über ihren Vor- und Nachnamen einem Prüfungsergebnis zugeordnet werden. Das entspricht 97.1 Prozent der Urliste. In einem nächsten Schritt wurde verglichen, ob sich Jugendliche, die an der Onlineumfrage teilgenommen haben, von Jugendlichen unterschieden, die auf die Einladung nicht reagiert haben. Bezüglich des Alters konnten keine Selektivitätsanalysen bestimmt werden, da es zu viele unplausible Geburtsdaten gab, die sich nicht ohne Weiteres korrigieren liessen. Im Hinblick auf das Geschlecht zeigte sich, dass unter den Teilnehmenden 54.0 Prozent und unter den Nichtteilnehmenden 49.2 Prozent Mädchen waren. Mädchen haben also eher an der Onlineumfrage teilgenommen als Jungen und dieser Unterschied ist statistisch signifikant, $F(1, 12'392) = 24.16$, $p < .001$. Insbesondere zeigten sich Unterschiede in den Prüfungsnoten. Teilnehmenden zeigten im Vergleich zu den Nichtteilnehmenden sowohl in Mathematik, $F(1, 12'392) = 103.95$, $p < .001$, als auch in Deutsch, $F(1, 12'392) = 101.94$, $p < .001$, und damit auch im Gesamtergebnis, $F(1, 12'392) = 142.89$, $p < .001$, bessere Prüfungsleistungen. Der Unterschied beträgt ungefähr ein Fünftel eines Notenpunkts und ist grösser für Mathematik als für Deutsch (Mathematik: $M_1 = 4.08$, $SD_1 = 1.15$ vs. $M_0 = 3.83$, $SD_0 = 1.29$; Deutsch: $M_1 = 4.31$, $SD_1 = 0.82$ vs. $M_0 = 4.14$, $SD_0 = 0.82$; Gesamt: $M_1 = 4.20$, $SD_1 = 0.83$ vs. $M_0 = 3.99$, $SD_0 = 0.90$). In **Tabelle 5** sind die jeweiligen Prüfungsnoten dargestellt.

Tabelle 5

Notenmittelwerte und Standardabweichungen der Teilnehmenden und Nichtteilnehmenden an der Onlinebefragung, getrennt nach Mathematik, Deutsch und Gesamt

Mittelwerte und Standardabweichungen der Prüfungsnoten von Teilnehmern und Nicht-Teilnehmern an der Onlinebefragung

	Mathematik		Deutsch		Gesamt	
Teilnahme	M	SD	M	SD	M	SD
ja	4.08	1.15	4.31	0.82	4.20	0.83
nein	3.83	1.29	4.14	0.82	3.99	0.90

Unter den erhobenen Merkmalen und Angaben gibt es zwei Konstrukte, die sich zu Skalen zusammenfassen lassen, und zwar (a) das emotionale Erleben während der Prüfungssituation (PANAS; Watson et al., 1988) sowie (b) Kontrollstrategien bei der Prüfungsvorbereitung (OPS; Tomasik et al., 2010).

Der PANAS erfasst, wie stark Personen in der Prüfungssituation positive bzw. negative Gefühle erleben. Dafür werden mehrere kurze Emotionsadjektive vorgegeben, zu denen die Befragten angeben, in welchem Ausmass sie die jeweiligen Emotionen während der

Prüfung empfunden haben. Aus den Angaben werden getrennte Kennwerte für positiven Affekt und negativen Affekt gebildet.

Der OPS erfasst, wie Jugendliche ihre Zielverfolgung und Selbstregulation in der Prüfungsvorbereitung steuern, insbesondere wenn Anforderungen hoch sind oder Schwierigkeiten auftreten. Im Fokus steht, ob sie sich aktiv engagieren (z. B. zielgerichtet weiterarbeiten, Anstrengung und Strategieeinsatz aufrechterhalten bzw. anpassen und sich bei Bedarf Unterstützung organisieren) oder ob sie disengagieren, indem sie sich selbstschützend von drohenden Misserfolgsimplikationen abgrenzen und/oder sich vom Ziel lösen (Zielablösung), um Belastung zu reduzieren (in Anlehnung an Heckhausen et al., 2010). Damit bildet die Skala zentrale Formen der motivationalen und volitionalen Steuerung in herausfordernden Vorbereitungssituationen ab.

Für die PANAS-Kurzversion bestätigt eine konfirmatorische Faktorenanalyse die erwartete zweifaktorielle Struktur (positiver vs. negativer Affekt), $\chi^2(8) = 51.58$, $p < .001$, CFI = .986, TLI = .975, RMSEA = .039. Positiver und negativer Affekt korrelieren auf latenter Ebene negativ ($r = -.28$), was mit Befunden aus der Literatur übereinstimmt. Für den OPS wird die fünffaktorielle Struktur durch eine konfirmatorische Faktorenanalyse bestätigt, $\chi^2(24) = 183.70$, $p < .001$, CFI = .967, TLI = .937, RMSEA = .043 (unter Zulassung einer Fehlerkorrelation). Die drei engagementbezogenen Strategiebereiche korrelieren untereinander positiv ($.48 < r < .66$), ebenso die beiden disengagementbezogenen Aspekte (Selbstprotektion und Zielablösung; $r = .41$). Korrelationen zwischen Engagement- und Disengagementaspekten fallen erwartungsgemäss gering aus und reichen von leicht negativ bis leicht positiv ($-.33 < r < .28$).

Zusammenfassend weisen sowohl der PANAS als auch der OPS sehr gute psychometrische Eigenschaften auf und die empirischen Korrelationsmuster stimmen weitgehend mit den theoretischen Erwartungen überein. Diese Befunde sprechen insgesamt für eine sehr gute Qualität der Fragebogendaten, und zwar auch, wo diese nicht faktoranalytisch getestet werden konnte.

3. Erleben der Prüfungssituation

Das Erleben der Prüfungssituation wurde entlang von drei Dimensionen erfasst. Zunächst konnten die Jugendlichen angeben, wie sie die Prüfung kognitiv bewerten. Ausgehend von einem Stress-Bewältigungs-Modell wurden hier die Aspekte der Herausforderung, Bedrohung, Kontrolle und Wichtigkeit erfasst. Danach wurde das emotionale Erleben mit zwei unterschiedlichen Skalen erfasst, die auf zwei unterschiedlichen theoretischen Konzeptionen beruhen. Schliesslich wurden die Jugendlichen gefragt, wie hoch sie ihre Bestehenswahrscheinlichkeit einschätzen, und zwar nachdem sie die Prüfung geschrieben haben, aber bevor sie das Prüfungsergebnis kennen. Die Ergebnisse zu den drei Dimensionen werden in den folgenden Abschnitten vorgestellt.

3.1. Kognitive Bewertung

Mit jeweils einem Item wurde erfasst, ob die Jugendlichen die Prüfung als Herausforderung («Ich habe die Prüfung als eine Herausforderung erlebt.») oder als Bedrohung («Ich habe die Prüfung als eine Bedrohung erlebt.») wahrnehmen, wie wichtig es ihnen war, die Prüfung zu bestehen («Es war mir sehr wichtig, die Prüfung zu bestehen.») und wie viel eigene Kontrolle sie über das Prüfungsergebnis hatten («Es hing vor allem von mir ab, ob ich die Prüfung bestehen kann.»).

Die Formulierung der Items erfolgte in Anlehnung an Gaab (2009), die zugrundeliegende theoretische Konzeption stammt von Lazarus und Folkman (1984). Demnach hängt die eigene emotionale Reaktion auf eine Situation von der primären und sekundären Bewertung dieser Situation bzw. der eigenen Bewältigungsressourcen ab. Das Resultat dieser Bewertung ist dann eine Einschätzung der Situation als bedrohlich, herausfordernd oder aber als irrelevant. Über verschiedene Anwendungsbereiche hinweg korreliert die Bewertung als Herausforderung mit besseren Bewältigungsergebnissen als die Bewertung als Bedrohung. Die Mittelwerte und Standardabweichungen der vier Items zum kognitiven Erleben finden sich in **Tabelle 6**.

Tabelle 6

Mittelwerte und Standardabweichungen nach den einzelnen Prüfungstypen und der kognitiven Bewertung

Mittelwerte und Standardabweichungen der Fragen zur kognitiven Bewertung

	Herausforderung		Bedrohung		Wichtigkeit		Kontrolle	
Prüfungstyp	M	SD	M	SD	M	SD	M	SD
LG	3.68	1.01	1.36	0.68	3.90	1.06	4.12	0.98
KG	3.79	0.97	1.71	0.91	4.12	1.01	4.13	0.95
MS	3.72	0.98	1.97	1.05	4.38	0.90	4.12	0.95
BM	3.56	0.94	1.55	0.83	3.94	1.06	4.16	0.91

Wenn man bedenkt, dass die Antworten auf einer 5-Punkte-Skala abgegeben worden sind, dann fällt auf, dass die Mittelwerte für Herausforderung, Wichtigkeit und vor allem Kontrolle relativ hoch sind, während die Mittelwerte für Bedrohung relativ niedrig ausgeprägt sind. Aus der Stress-Bewältigungs-Forschung (z. B. Folkman et al., 1986) ist bekannt, dass diese Art von kognitiver Bewertung mit einem problemorientierten Bewältigungsstil einhergeht und damit in der Regel auch mit positiveren Bewältigungsergebnissen.

Dieses erwartbare Muster zeigt sich teilweise auch in den Daten aus dieser Onlinebefragung, in der auch die Strategien zur Prüfungsvorbereitung nach Heckhausen et al. (2010; vgl. auch Tomasik et al., 2010) erfragt wurden. So zeigt sich, dass es insbesondere die Einschätzung der Wichtigkeit ist, die mit den Strategien zur Prüfungsvorbereitung korreliert ist. Jugendliche, die die Prüfung als wichtiger

einschätzen, berichten auch, dass sie mehr Zeit und Energie in die Prüfungsvorbereitung investieren (r = .25), mehr motivationale Strategien verwenden (r = .41), häufiger Hilfe und Unterstützung suchen (r = .20) und weniger bereit sind, sich mit einem Misserfolg abzufinden (r = -.41). Ausserdem korreliert Valenz negativ mit Herausforderung (r = −.20) und besonders mit Bedrohung (r = −.39); Erregung korreliert positiv mit Herausforderung (r = .22), Bedrohung (r = .27) und Wichtigkeit (r = .21); Kontrollerleben korreliert negativ mit Bedrohung (r = −.26); Bedrohung korreliert mit weniger positivem Affekt (r = −.28) und mehr negativem Affekt (r = .43). Alle anderen Aspekte der kognitiven Bewertung korrelieren so gut wie gar nicht mit den Strategien zur Prüfungsvorbereitung. Eine Wirkungsrichtung lässt sich aus diesen korrelativen Befunden allerdings nicht ableiten. Es kann sein, dass Jugendliche, denen die Prüfung wichtiger war, sich aus diesem Grund auch effektiver vorbereitet haben. Es kann allerdings auch sein, dass eine effektivere Prüfungsvorbereitung dazu geführt hat, dass sie die Prüfung als wichtiger wahrnehmen.

Hinsichtlich der Einschätzung der Prüfung als Herausforderung unterscheiden sich die Prüfungstypen signifikant, $F(3, 3'597) = 6.12$, $p = .001$, wobei sich im Einzelvergleich herausstellt, dass sich die KG- und BM-Prüfungen sowie die KG- und LG-Prüfungen signifikant unterscheiden. Die Prüfung für das KG wird also als besonders herausfordernd erlebt.

Grössere Unterschiede gibt es bei der Einschätzung der Prüfung als Bedrohung. Hier ist nicht nur die Varianzanalyse über alles signifikant, $F(3, 3'597) = 68.30$, $p < .001$, sondern alle Prüfungstypen unterscheiden sich im Einzelvergleich voneinander. Die Einschätzung als Bedrohung ist bei den MS-Prüfungen am höchsten und bei den LG-Prüfungen am niedrigsten. Zudem ist die Streuung bei den MS-Prüfungen ziemlich hoch, was darauf hindeutet, dass es nur ein Teil der Jugendlichen ist, der hier zu so einer Einschätzung kommt.

Auch bei der Einschätzung der Wichtigkeit unterscheiden sich die Prüfungstypen signifikant, $F(3, 3'597) = 21.19$, $p < .001$, wobei es im Einzelvergleich lediglich zwischen der LG-Prüfung und der BM-Prüfung keine Unterschiede gibt. Insgesamt wird die MS-Prüfung als besonders wichtig eingeschätzt und hier gibt es auch eine relativ geringe Streuung. Bezüglich des Erlebens der eigenen Kontrolle gibt es keine signifikanten Unterschiede zwischen den Prüfungstypen, $F(3, 3'597) = 0.23$, $p = 0.88$.

Zusammenfassend zeigt sich, dass die Prüfung von den Jugendlichen im Mittel deutlich stärker als Herausforderung denn als Bedrohung eingeschätzt wird. Gleichzeitig wird sie als sehr wichtig erlebt und es wird vergleichsweise über viel eigene Kontrolle berichtet. Auf der verwendeten 5-Punkte-Skala liegen diese Einschätzungen damit insgesamt eher auf der «hohen» Seite (Herausforderung/Wichtigkeit/Kontrolle) beziehungsweise eher auf der «tiefen» Seite (Bedrohung). In der Stress- und Bewältigungsforschung wird dieses Muster typischerweise als Hinweis darauf gelesen, dass eine Situation zwar als bedeutsam, aber grundsätzlich als bewältigbar wahrgenommen wird, was eher mit günstigeren Bewältigungsprozessen einhergeht als eine dominant bedrohliche Bewertung (Lazarus & Folkman, 1984; Folkman et al., 1986). Für die Interpretation ist

zusätzlich wichtig, dass hier pro Aspekt jeweils nur ein Item verwendet wurde: Die Mittelwerte sind damit gut als grobe Orientierungswerte für die Selbstwahrnehmung zu verstehen, weniger als präzise Messung stabiler Dispositionen. Auffällig sind innerhalb dieses insgesamt erwartbaren Befundmusters die MS-Prüfungen, da sie im Vergleich als besonders wichtig und zugleich als relativ bedrohlich erlebt werden.

3.2. Emotionales Erleben

Das emotionale Erleben während der Prüfungssituation wurde mit der PANAS-Kurzskala erfasst, deren Skalenbildung weiter oben beschrieben worden ist. Darüber hinaus wurde der Self-Assessment-Manikin (SAM; Bradley & Lang, 1994) eingesetzt, mit dem die Aspekte der Valenz, der Erregung und der Kontrolle erfasst worden sind. Die zwei Operationalisierungen entstammen unterschiedlichen theoretischen Traditionen. Beim PANAS kommen zur Einschätzung des eigenen emotionalen Erlebens Adjektive zum Einsatz, die positive bzw. negative Emotionen beschreiben. Dabei wurden die Adjektive so ausgewählt, dass die positiven bzw. negativen untereinander hoch korrelieren, aber mit den jeweils anderen Dimensionen (weitgehend) unkorreliert sind. Folglich resultieren üblicherweise zwei statistisch (weitgehend) unabhängige Dimensionen des emotionalen Erlebens. Beim SAM geht man im sogenannten dimensionalen Ansatz davon aus, dass sich alle Emotionen auf einigen wenigen Dimensionen abbilden lassen, wovon die Valenz und die Erregung die wichtigsten sind.

Insgesamt ist der positive Affekt leicht überdurchschnittlich und der negative Affekt leicht unterdurchschnittlich ausgeprägt, was sich im SAM in einer tendenziell positiven Valenz widerspiegelt. Die Erregung wird als leicht unterdurchschnittlich bis durchschnittlich beschrieben und die wahrgenommene Kontrolle über die Prüfungssituation ist durchschnittlich bis leicht überdurchschnittlich. In **Tabelle 7** finden sich die Mittelwerte und Standardabweichungen der Skalen bzw. Items.

Tabelle 7

Mittelwerte und Standardabweichungen nach den einzelnen Prüfungstypen und der emotionalen Bewertung

Mittelwerte und Standardabweichungen der Fragen zur kognitiven Bewertung

	Positiver Affekt		Negativer Affekt		Valenz		Erregung		Kontrolle	
Prüfungstyp	M	SD	M	SD	M	SD	M	SD	M	SD
LG	3.69	0.66	1.87	0.67	3.82	0.73	2.64	1.02	3.71	0.74
KG	3.41	0.70	2.45	0.86	3.37	0.83	2.94	1.11	3.32	0.82
MS	3.24	0.69	2.50	0.84	3.18	0.83	3.07	1.08	3.26	0.84
BM	3.32	0.65	2.32	0.85	3.43	0.76	2.87	1.14	3.28	0.81

Das emotionale Erleben unterscheidet sich zuweilen deutlich zwischen den Prüfungstypen, wobei hier die LG-Prüfung heraussticht. Bezüglich des positiven Affekts unterscheiden sich die Prüfungstypen signifikant, $F(3, 3'597) = 71.20$, $p < .001$. Im

Einzelvergleich unterscheidet sich die LG-Prüfung von allen anderen drei Prüfungstypen und zeichnet sich durch ein deutlich positiveres emotionales Erleben aus. Ausserdem unterscheidet sich die KG-Prüfung von der MS-Prüfung.

Auch bezüglich des negativen Affekts gibt es signifikante Unterschiede zwischen den Prüfungstypen, $F(3, 3'597) = 157.06$, $p < .001$. Auch hier unterscheidet sich die LG-Prüfung von allen anderen drei Prüfungstypen und zwar insofern, als dass sich die LG-Prüfung durch den geringsten negativen Affekt auszeichnet. Ausserdem unterscheidet sich die BM-Prüfung von allen anderen drei Prüfungstypen und liegt somit zwischen der LG-Prüfung einerseits und der KG- bzw. der MS-Prüfung andererseits.

Auch die Valenz unterscheidet sich signifikant zwischen den Prüfungstypen, $F(3, 3'597) = 112.34$, $p < .001$. Im Einzelvergleich zeigt sich wieder, dass sich die LG-Prüfung von allen anderen drei Prüfungstypen unterscheidet. Ausserdem gibt es signifikante Unterschiede zwischen der KG-Prüfung und der MS-Prüfung sowie zwischen der BM-Prüfung und der MS-Prüfung.

Signifikante Unterschiede gibt es auch bezüglich der Erregung, $F(3, 3'597) = 24.63$, $p < .001$, wobei sich hier im Einzelvergleich die LG-Prüfung von allen anderen drei Prüfungstypen unterscheidet. Das gleiche Bild zeigt sich bezüglich der wahrgenommenen Kontrolle. Bei der LG-Prüfung wird also sowohl weniger Erregung als auch mehr Kontrolle über die Prüfungssituation erlebt, wenn man das mit den anderen drei Prüfungstypen vergleicht.

Auch das emotionale Erleben während der Prüfungssituation korreliert fast gar nicht mit den Prüfungsnoten. Lediglich der positive Affekt ($r = .18$) und die wahrgenommene Kontrolle ($r = .29$) korrelieren positiv mit der Mathematiknote, interessanterweise aber nicht mit der Deutschnote.

In der Gesamtbetrachtung ergibt sich ein stimmiges Bild über beide Erfassungszugänge hinweg: Der positive Affekt ist eher höher und der negative Affekt eher tiefer ausgeprägt, was sich zugleich in einer tendenziell positiven Valenz widerspiegelt; Erregung liegt eher im mittleren Bereich, und die wahrgenommene Kontrolle über die Prüfungssituation ist durchschnittlich bis leicht erhöht. Das ist insofern gut einzuordnen, als die Prüfung gleichzeitig als wichtig und eher als Herausforderung erlebt wird. Unterschiede zwischen den Prüfungstypen sind meist nicht sehr gross, aber punktuell klar erkennbar: Besonders die LG-Prüfung sticht durchgängig durch ein vorteilhafteres emotionales Erleben heraus (mehr positiver Affekt, weniger negativer Affekt, positivere Valenz, weniger Erregung und mehr Kontrolle). Für das Lesen der Befunde hilft ausserdem eine begriffliche Klarstellung: «Kontrolle» erscheint hier in zwei Bedeutungen – als Teil der emotionalen Einschätzung (SAM-Kontrollurteil) und als kognitives Urteil zur eigenen Einflussmöglichkeit (Item aus 3.1). Beide sind inhaltlich verwandt, aber nicht identisch, was erklärt, warum sie sich in den Korrelationsmustern unterschiedlich verhalten können.

3.3. Subjektive Bestehenswahrscheinlichkeit

Die Jugendlichen wurden gebeten, anzugeben, wie wahrscheinlich sie es gerade finden, dass sie die Prüfung bestanden haben werden. Diese Wahrscheinlichkeit wurde auf einer Skala von 0 bis 100 Prozent abgefragt, wobei zum Zeitpunkt der Befragung keiner der Jugendlichen das offizielle Prüfungsergebnis schon wissen konnte. Der Modalwert der Verteilung liegt für alle Prüfungstypen bei 50 Prozent. Auch der Mittelwert liegt um die 50 Prozent mit Ausnahme der LG-Prüfung, bei der die Jugendlichen eine etwas positivere Einschätzung ihrer Chancen haben. Die Mittelwerte und Standardabweichungen der subjektiven Bestehenswahrscheinlichkeit finden sich in **Tabelle 8**.

Eine höhere subjektive Bestehenswahrscheinlichkeit korreliert mit mehr positivem Affekt (r = .46), weniger negativem Affekt (r = -.35), einer positiveren Valenz (r = .42), weniger Erregung (r = -.16) und vor allem einer höheren wahrgenommenen Kontrolle (r = .53). Alle Korrelationen entsprechen bezüglich ihres Vorzeichens und ihrer Grössenordnung den theoretischen Erwartungen. Die Korrelationen mit der kognitiven Bewertung sind deutlich niedriger und betragen für Herausforderung r = -.14, für Bedrohung r = -.18, für Wichtigkeit r = .27 und für Kontrolle r = .06. Gleichzeitig zeigt sich, dass die Jugendlichen ihre Bestehenswahrscheinlichkeit einigermassen zutreffend einschätzen können. Es zeigen sich positive Korrelationen mit der Mathematiknote (r = .41), der Deutschnote (r = .21) sowie der Gesamtnote (r = .39).

Tabelle 8

Mittelwerte und Standardabweichungen der subjektiven Bestehenswahrscheinlichkeit, aufgeteilt nach den einzelnen Prüfungstypen

Mittelwerte und Standardabweichungen der subjektiven Bestehenswahrscheinlichkeit

Prüfungstyp	M	SD
LG	61.08	18.86
KG	51.92	21.11
MS	51.92	21.42
BM	52.88	21.73

Zusammenfassend deutet die Verteilung der subjektiven Bestehenswahrscheinlichkeit (Modalwert 50 Prozent in allen Prüfungstypen, Mittelwerte insgesamt um 50 Prozent mit leicht höherer Einschätzung bei der LG-Prüfung) darauf hin, dass viele Jugendliche unmittelbar nach der Prüfung zwar eine Tendenz spüren, aber gleichzeitig merkliche Unsicherheit berichten. Die Korrelationsmuster helfen beim Einordnen: Die subjektive Bestehenswahrscheinlichkeit hängt deutlich stärker mit dem emotionalen Erleben und insbesondere mit situativ erlebter Kontrolle zusammen als mit den kognitiven

Bewertungen «Herausforderung/Bedrohung». Das spricht dafür, dass Jugendliche ihre Einschätzung stark aus unmittelbaren affektiven und kontrollebenen Hinweisreizen ableiten (wie «ich hatte es im Griff»), weniger aus der abstrakteren Einordnung der Prüfung als Herausforderung oder Bedrohung. Gleichzeitig zeigen die positiven Zusammenhänge mit den Prüfungsnoten (vor allem Mathematik und Gesamtnote), dass die Einschätzung im Mittel nicht beliebig ist, sondern in einem relevanten Ausmass die tatsächliche Leistung widerspiegelt. Wichtig bleibt dabei: Es handelt sich um eine Selbsteinschätzung nach der Prüfung, also um eine Einschätzung auf Basis eigener Eindrücke – sie ist deshalb als subjektive «Kalibrierung» der eigenen Performance zu lesen, nicht als unabhängige Prognose.

4. Besuch von Vorbereitungskursen

Die bisher dargestellten Befunde beschreiben das Erleben der Prüfungssituation. Davon zu unterscheiden ist die Frage, wie sich die Jugendlichen konkret auf die Prüfung vorbereitet haben. Um der Frage nach der Nutzung von öffentlichen bzw. privaten Vorbereitungskursen nachzugehen, wurden zwei Aussagen formuliert, die auf die Regelmässigkeit der Nutzung abgezielt haben («Um mich für die Prüfung vorzubereiten, besuchte ich regelmässig einen von der Schule angebotenen Vorbereitungskurs.» und «Um mich für die Prüfung vorzubereiten, habe ich regelmässig einen privaten Vorbereitungskurs besucht.»). Die Begriffe «öffentlich» und «schulisch» werden in diesem Bericht synonym verwendet. Die Jugendlichen konnten den Aussagen auf einer 5-Punkte-Skala zustimmen.

Wie **Abbildung 14** zur Nutzung von öffentlichen Vorbereitungskursen zu entnehmen ist, zeigte sich bei allen Prüfungstypen eine bimodale Verteilung, die darauf schliessen lässt, dass ein Kurs entweder regelmässig oder gar nicht besucht worden ist. Eine sehr ähnliche Verteilung zeigt sich auch bei der Nutzung von privaten Vorbereitungskursen.

Aus diesem Grund wurden für die folgenden Auswertungen beide Variablen dichotomisiert und dabei einerseits die Kategorien 1 («stimmt gar nicht») und 2 («stimmt wenig») und andererseits die Kategorien 3 («stimmt mittel»), 4 («stimmt ziemlich») sowie 5 («stimmt voll») zusammengefasst (Altman & Royston, 2006; MacCallum et al., 2002). Gemäss dieser Operationalisierung nutzen 68.8 Prozent der Jugendlichen einen öffentlichen und 40.6 Prozent einen privaten Vorbereitungskurs mit zumindest mittelmässiger Regelmässigkeit (vgl. Altman & Royston, 2006; MacCallum et al., 2002).

Fast alle befragten Schülerinnen und Schüler besuchten einen Vorbereitungskurs für die ZAP. Rund zwei Drittel oder 68.8 Prozent besuchten einen öffentlichen Vorbereitungskurs. Einen privaten Vorbereitungskurs besuchte knapp die Hälfte oder 40.6 Prozent der Schülerinnen und Schüler. Immerhin 18.2 Prozent besuchten sowohl einen öffentlichen als auch einen privaten Vorbereitungskurs und nur 8.8 Prozent verzichteten ganz auf einen Besuch.

Abbildung 14

Anzahl der Nutzung öffentlicher Vorbereitungskurse nach den einzelnen Prüfungstypen

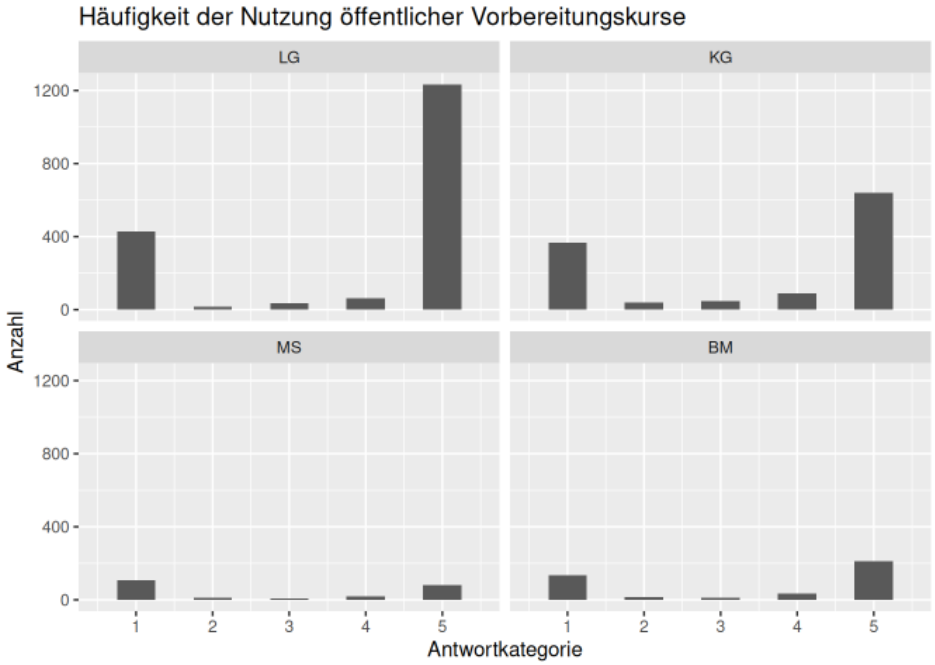
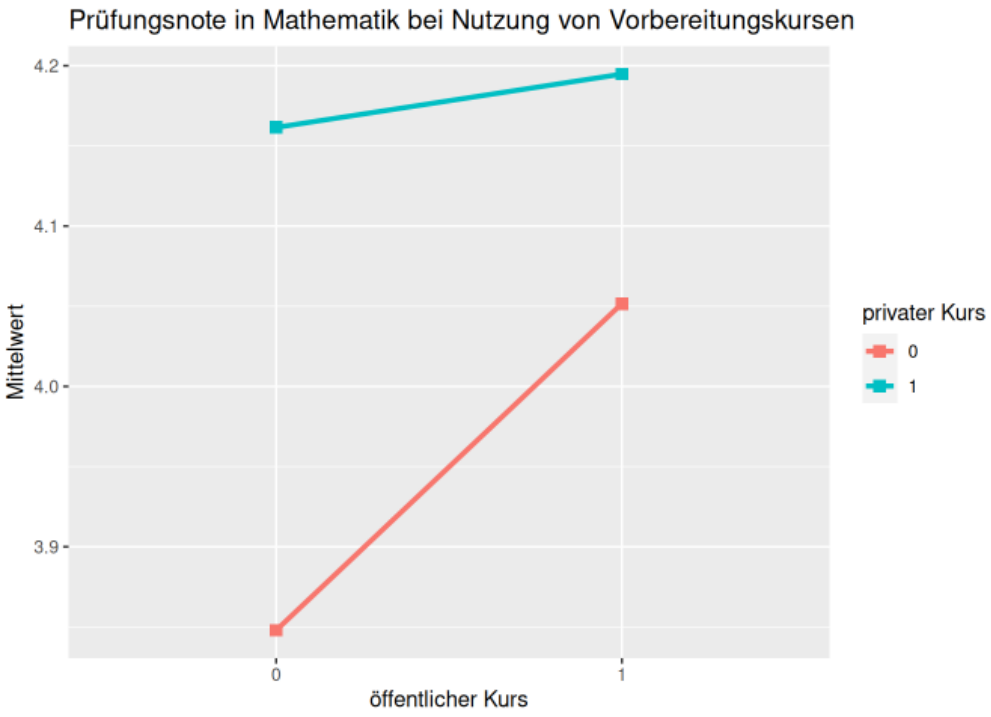


Abbildung 15

Mittelwert der Mathematiknote vor und nach der Prüfung bei Nutzung von einem privaten vs. öffentlichen Vorbereitungskurs



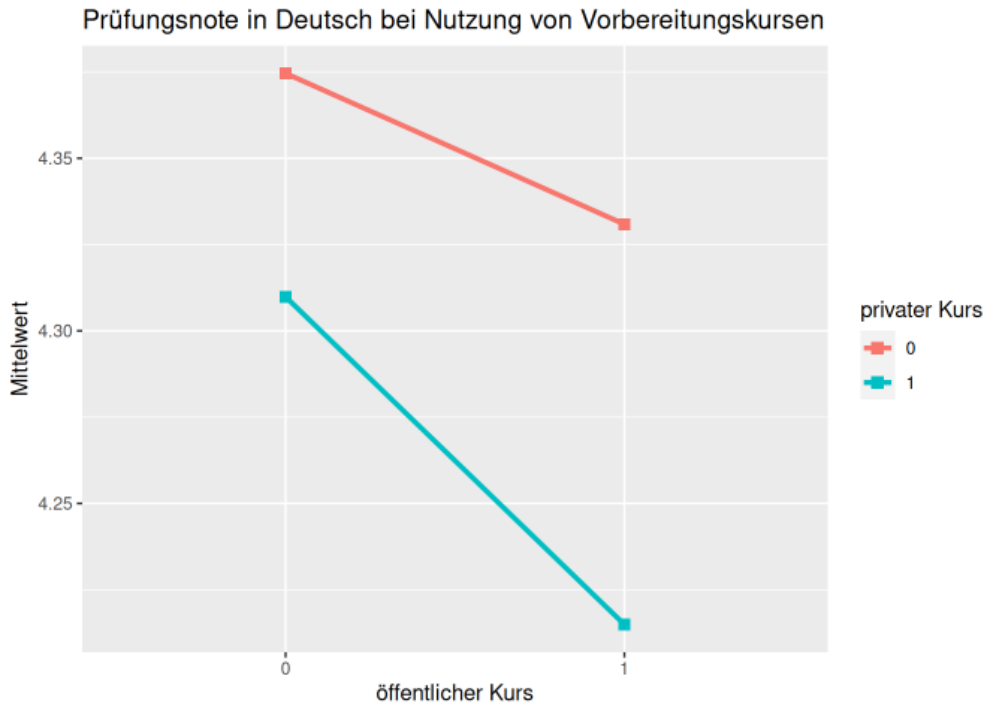
Der Besuch eines Vorbereitungskurses ging mit besseren Prüfungsnoten im Fach Mathematik einher. Wer einen privaten Kurs besucht hatte war unabhängig vom Besuch

weiterer Kurse 0.23 Notenpunkte besser als Schülerinnen und Schüler, die keinen privaten Kurs besucht hatten, $F(1, 3492) = 20.56$, $p < 0.001$. Wer hingegen einen öffentlichen Kurs besucht hatte war unabhängig vom Besuch weiterer Kurse 0.12 Notenpunkte besser als Schülerinnen und Schüler, die keinen öffentlichen Kurs besucht hatten, $F(1, 3496) = 5.30$, $p < 0.05$. Dieser Zusammenhang ist in **Abbildung 15** dargestellt.

Für das Fach Deutsch fand sich ein statistischer Zusammenhang, der sich jedoch in einem negativen Zusammenhang mit der Prüfungsnote äusserte. Allerdings war dieser so gering, dass er in der Praxis wenig bedeutsam ist. Besucher des privaten Kurses waren unabhängig vom Besuch weiterer Kurse 0.09 Notenpunkte schlechter als solche, die keinen privaten Kurs besucht hatte, $F(1, 3492) = 10.22$, $p < 0.001$). Besucher des öffentlichen Kurses waren unabhängig vom Besuch weiterer Kurse 0.07 Notenpunkte schlechter als solche, die keinen privaten Kurs besucht hatten, $F(1, 3492) = 4.882$, $p < 0.05$). Dieser Zusammenhang ist in **Abbildung 16** dargestellt. Für die Gesamtprüfungsleistung mitteln sich die Unterschiede bei den Mathematik- bzw. Deutschnoten aus, sodass sich am Ende kein Zusammenhang zwischen Prüfungsleistung und Besuch eines privaten oder öffentlichen Vorbereitungskurses zeigt.

Abbildung 16

Mittelwert der Deutschnote vor und nach der Prüfung bei Nutzung von einem privaten vs. öffentlichen Vorbereitungskurs



Für die Einordnung der Befunde ist zunächst wichtig zu berücksichtigen, dass die Nutzung öffentlicher und privater Vorbereitungskurse in den Daten nicht als fein abgestufte Intensität erscheint, sondern faktisch als «entweder regelmässig oder gar nicht». Entsprechend ist die gewählte Dichotomisierung (1–2 vs. 3–5) vor allem als

pragmatische Abbildung dieser beobachteten Zweiteilung zu verstehen: Sie reduziert Komplexität, ohne ein eigentlich kontinuierliches Muster künstlich zu «zerhacken». Gleichzeitig bleibt damit die genaue Intensität innerhalb der Gruppe der regelmässigen Nutzer*innen (beispielsweise «mittel» vs. «voll») unsichtbar, weshalb die Ergebnisse als Unterschiede zwischen zwei groben Nutzungsprofilen zu lesen sind und nicht als Dosis-Wirkungs-Beziehung (Altman & Royston, 2006; MacCallum et al., 2002).

Die Zusammenhänge zwischen Kursbesuch und Prüfungsleistung sind zudem konsequent als *korrelative* Befunde zu interpretieren. Dass sich beim privaten Vorbereitungskurs in Mathematik ein positiver, in Deutsch jedoch ein negativer Zusammenhang zeigt, während sich für die Gesamtleistung kein Zusammenhang ergibt, passt zu einem Bild punktueller, fachbezogener Zusammenhänge, die sich auf Gesamtniveau teilweise gegenseitig aufheben. Aus diesen Mustern lässt sich jedoch nicht ableiten, ob ein Vorbereitungskurs die Leistung verbessert oder verschlechtert: Selbstselektion bleibt eine naheliegende Erklärung, da sich Kursnutzer*innen und Nichtnutzer*innen in nicht beobachteten Merkmalen unterscheiden können (etwa Lernstrategien, Prüfungsangst, Unterstützungsnetzwerk oder wahrgenommene Passung zwischen Kurs und Prüfungsanforderungen). Dass sich zumindest dort, wo die Vornote in Deutsch vorliegt, kein Zusammenhang zwischen Vornote und Kursbesuch zeigt, $F(1, 2'953) = 2.56$, $p = .11$, ist ein hilfreicher Hinweis gegen eine ganz einfache Selektionsgeschichte – ersetzt aber keine kausale Identifikation, insbesondere weil weiterhin Drittvariablen verantwortlich sein können (Holland, 1986).

5. Weitere institutionelle Unterstützung

Neben dem Besuch von öffentlichen oder privaten Vorbereitungskursen wurde auch abgefragt, ob die Jugendlichen von der Schule («Meine Schule hat mich sehr dabei unterstützt, mich auf die Prüfung vorzubereiten.») oder von den Eltern («Meine Eltern haben mich sehr dabei unterstützt, mich auf die Prüfung vorzubereiten.») bei der Prüfungsvorbereitung unterstützt wurden und inwieweit sie sich an einem Informationsanlass über die Prüfung informieren konnten («Ich konnte mich an einem Informationsabend sehr gut über die Prüfung informieren.»). Durch die Formulierung wurde ein hohes Mass an Unterstützung abgefragt. Die Antwortskalen reichten jeweils von 1 bis 5. Die Mittelwerte der Antworten finden sich in **Tabelle 9**.

Wenn man sich neben den Mittelwerten und Standardabweichungen auch die Verteilung der wahrgenommenen Unterstützung durch die Schule ansieht, zeigt sich eine hohe Varianz in den Einschätzungen. Diese ist in **Abbildung 17** zu sehen. Obwohl es viele Jugendliche gibt, die die Unterstützung durch die Schule als hoch einschätzen, gibt es einen fast ebenso grossen Anteil, der die Unterstützung als nicht gegeben sieht. Das gilt insbesondere für Jugendliche, die sich auf eine Mittelschulprüfung angemeldet haben.

Besonders augenfällig ist diese Varianz der Verteilung, wenn man sie mit der Varianz für die Einschätzung der Unterstützung durch die Eltern vergleicht. Diese ist in **Abbildung 18**

dargestellt. Hier sieht man, dass sich die allermeisten Jugendlichen sehr gut durch ihre Eltern unterstützt gefühlt haben. Das gilt insbesondere für die Jugendlichen, die sich für die LG-Prüfung angemeldet haben. Auch der Nutzen von Informationsveranstaltungen wird sehr unterschiedlich eingeschätzt, was zum Teil daran liegen kann, dass diese gar nicht angeboten oder besucht worden sind. Die Verteilung ist in **Abbildung** zu sehen.

Tabelle 9

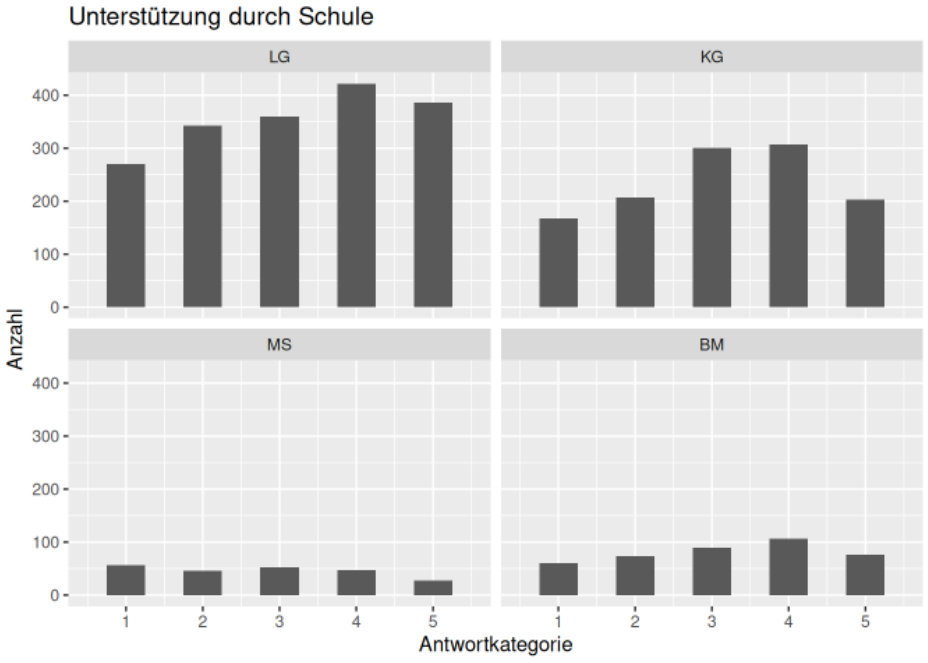
Mittelwerte und Standardabweichungen nach Prüfungstyp und Art sozialer Unterstützung

Mittelwerte und Standardabweichungen der Fragen zur Unterstütrung

	Schule		Eltern		Infoanlass	
Prüfungstyp	M	SD	M	SD	M	SD
LG	3.17	1.37	4.59	0.76	2.87	1.49
KG	3.15	1.29	4.15	1.13	2.47	1.44
MS	2.75	1.34	3.98	1.19	2.36	1.45
BM	3.15	1.33	3.79	1.28	1.75	1.21

Abbildung 17

Schulische Unterstützung nach den einzelnen Prüfungstypen und Antwortkategorien



Vergleicht man die Unterstützung zwischen den unterschiedlichen Prüfungstypen, dann zeigen sich für die Unterstützung durch die Schule signifikante Unterschiede, $F(3, 3'597) = 6.86$, $p < .001$, sodass sich die MS-Prüfung mit einer besonders tiefen Einschätzung signifikant von allen anderen Prüfungstypen unterscheidet. Die anderen Prüfungstypen wiederum unterscheiden sich in dieser Hinsicht nicht signifikant voneinander.

Auch bezüglich der Unterstützung durch die Eltern gibt es signifikante Unterschiede zwischen den Prüfungstypen, $F(3, 3'597) = 104.77$, $p < .001$. Im Einzelvergleich sieht man, dass sich die LG-Prüfung signifikant von allen anderen unterscheidet. Die höheren Werte bei der LG-Prüfung sind vielleicht darauf zurückzuführen, dass die Jugendlichen da deutlich jünger als bei den anderen Prüfungstypen sind und deswegen auch mehr Unterstützung bekommen oder einfordern. Ausserdem unterscheidet sich die KG-Prüfung von der BM-Prüfung.

Abbildung 18

Elterliche Unterstützung nach den einzelnen Prüfungstypen und Antwortkategorien

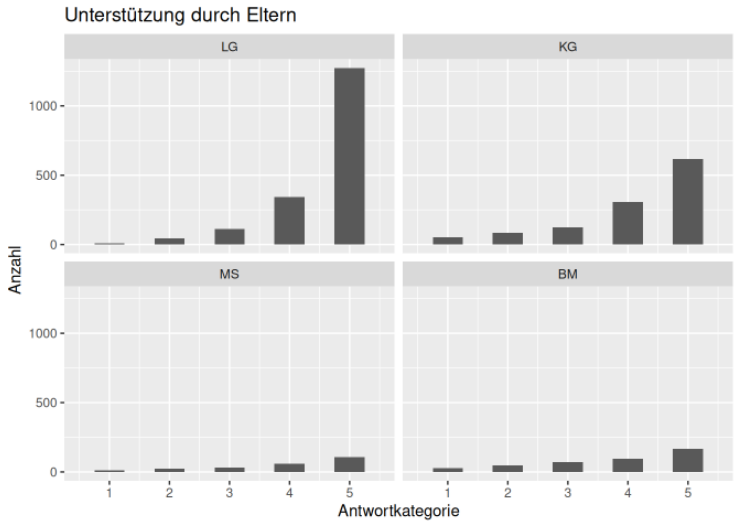
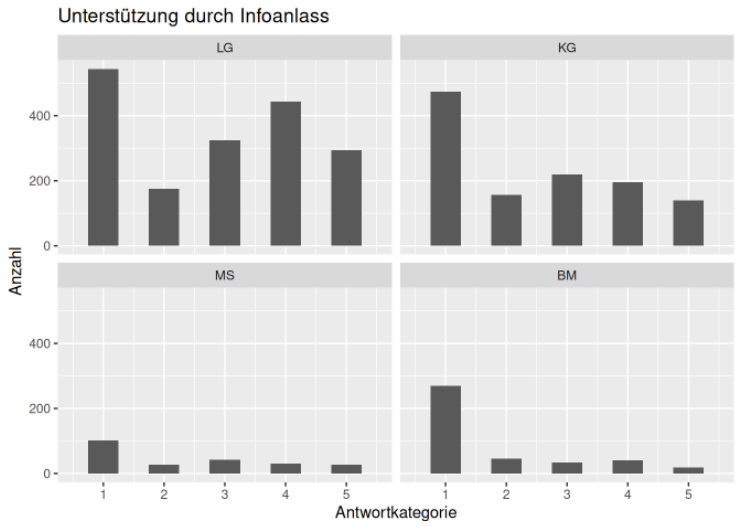


Abbildung 19

Nutzung von Informationsveranstaltungen nach den einzelnen Prüfungstypen und Antwortkategorien



Auch für die wahrgenommene Nützlichkeit einer Informationsveranstaltung zeigen sich gesamthaft signifikante Unterschiede zwischen den Prüfungstypen, $F(3, 3'597) = 73.36$, $p < .001$. Bis auf die KG-Prüfung und die MS-Prüfung, zwischen denen es keinen signifikanten Unterschied gibt, unterscheiden sich ansonsten alle Prüfungstypen voneinander. Wie in **Abbildung** ersichtlich, wird dabei die Informationsveranstaltung für die LG-Prüfung als am informativsten wahrgenommen, die für die BM-Prüfung als am wenigsten informativ.

Für die Einordnung der Befunde ist zunächst wichtig festzustellen, was hier genau gemessen wurde: Die Items sind bewusst so formuliert, dass sie ein *sehr hohes* Ausmass an Unterstützung erfassen («sehr unterstützt», «sehr gut informieren»). Mittelwerte im mittleren Bereich sind deshalb nicht automatisch «mittelmässige» Unterstützung, sondern können auch bedeuten, dass zwar Unterstützung vorhanden war, aber nicht in einem Ausmass, das die Jugendlichen als «sehr» einstufen würden. Für das Lesen der Verteilungen ist ausserdem hilfreich, nicht nur auf Mittelwerte zu schauen: Gerade bei der schulischen Unterstützung zeigt sich eine ausgeprägte Streuung bis hin zu einer Art Polarisierung (ein grosser Teil berichtet hohe Unterstützung, ein ebenfalls grosser Teil berichtet praktisch keine). Bei der Unterstützung durch die Eltern ist das Muster dagegen deutlich konzentrierter in Richtung hoher Zustimmung, während der wahrgenommene Nutzen von Informationsveranstaltungen auch deshalb stark variiert, weil solche nicht überall angeboten oder tatsächlich besucht worden sein dürften.

Dass die wahrgenommene Unterstützung praktisch nicht mit den Prüfungsnoten zusammenhängt (maximal kleine Zusammenhänge), ist vor diesem Hintergrund nicht überraschend und sollte nicht als «Unwirksamkeit» gelesen werden. Erstens sind es korrelative Selbstauskünfte und Unterstützung kann reaktiv sein (mehr Unterstützung gerade dann, wenn ein*e Schüler*in sich unsicher fühlt). Zweitens bildet eine globale Selbsteinschätzung («sehr unterstützt») nicht zwingend die fachlich passende Unterstützung ab, die sich direkt in einer Mathematik- oder Deutschnote niederschlagen würde. Drittens können Deckeneffekte (besonders bei elterlicher Unterstützung) die statistischen Zusammenhänge zusätzlich dämpfen, weil dann zwischen den meisten Jugendlichen kaum noch Unterschiede messbar sind.

6. Abdeckung des Prüfungsstoffs durch die Volksschule

Unabhängig von der Unterstützung stellt sich zudem die Frage, inwieweit der Prüfungsstoff aus Sicht der Jugendlichen durch Unterricht und Vorbereitung tatsächlich abgedeckt war. Mit jeweils zwei Fragen wurde nach der Abdeckung des Prüfungsstoffs in Mathematik und Deutsch durch den schulischen Unterricht bzw. durch einen an der Schule angebotenen (also öffentlichen) Vorbereitungskurs gefragt. Bei den Auswertungen bezüglich des Vorbereitungskurses wurden alle Jugendlichen ausgeschlossen, die

angaben, keinen solchen Kurs besucht zu haben. Die Mittelwerte und Standardabweichungen der Auswertungen zur Abdeckung des Prüfungsstoffs finden sich in **Tabelle** .

Tabelle 18

Mittelwerte und Standardabweichungen der Auswertungen zur Abdeckung des Prüfungsstoffs, aufgeteilt nach Prüfungstyp, Mathematik- und Deutsch-Unterricht vs. Mathematik- und Deutsch-Vorbereitungskurs

	Vorbereitungskurs Mathematik		Vorbereitungskurs Deutsch	
Prüfungstyp	M	SD	M	SD
LG	3.92	1.06	3.99	1.04
KG	3.57	1.18	3.68	1.16
MS	3.38	1.24	3.51	1.24
BM	3.54	1.23	3.59	1.24

	Unterricht Mathematik		Unterricht Deutsch	
Prüfungstyp	M	SD	M	SD
LG	2.34	1.10	2.77	1.09
KG	2.76	1.15	3.08	1.13
MS	2.72	1.11	3.18	1.12
BM	3.14	1.10	3.38	1.13

Die vier Prüfungstypen unterscheiden sich signifikant hinsichtlich der wahrgenommenen Abdeckung des Prüfungsstoffs im schulischen Mathematikunterricht, $F(3, 3'597) = 72.69$, $p < .001$. Im Einzelvergleich wird deutlich, dass sich alle Prüfungstypen voneinander unterscheiden. Die einzige Ausnahme bilden die KG- und die MS-Prüfung, die sich nicht signifikant voneinander unterscheiden.

Ein ähnliches Muster zeigt sich hinsichtlich der wahrgenommenen Abdeckung des Prüfungsstoffs im schulischen Deutschunterricht. Es zeigen sich signifikante Unterschiede zwischen den Prüfungstypen, $F(3, 3'597) = 43.53$, $p < .001$, und im Einzelvergleich unterscheidet sich die LG-Prüfung von allen anderen sowie die KG-Prüfung von der BM-Prüfung.

Die Abdeckung des Prüfungsstoffs durch den jeweiligen Unterricht wird als unzureichend bis mittelmässig angesehen. Ausserdem gilt für alle Prüfungstypen, dass die Einschätzungen für Mathematik zuweilen deutlich unter denen für Deutsch liegen. Die Bewertung des schulischen Vorbereitungskurses fällt dagegen deutlich positiver aus. Die Verteilungen für beide Fächer und beide Unterstützungsarten finden sich in **Abbildung 19** und **Abbildung 20**. Bezüglich der Abdeckung des Prüfungsstoffs durch einen schulischen Vorbereitungskurs zeigen sich signifikante Unterschiede sowohl für Mathematik, $F(3, 2'916) = 29.25$, $p < .001$, als auch für Deutsch, $F(3, 2'916) = 24.75$, $p < .001$, was in **Abbildung 21** und **Abbildung 22** zu sehen ist. In beiden Fällen sind diese Unterschiede darauf zurückzuführen, dass sich die LG-Prüfung von allen anderen Prüfungen unterscheidet.

Abbildung 19

Abdeckung des Mathematik-Prüfungsstoffs durch Unterricht nach den einzelnen Prüfungstypen und Antwortkategorien

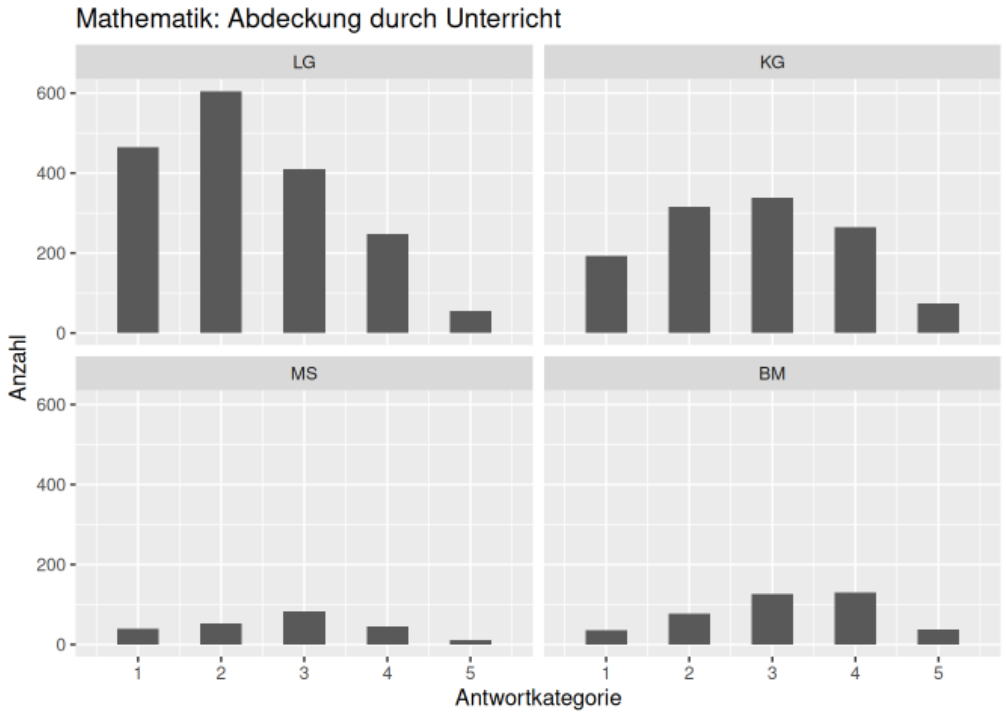


Abbildung 20

Abdeckung des Deutsch-Prüfungsstoffs durch Unterricht nach den einzelnen Prüfungstypen und Antwortkategorien

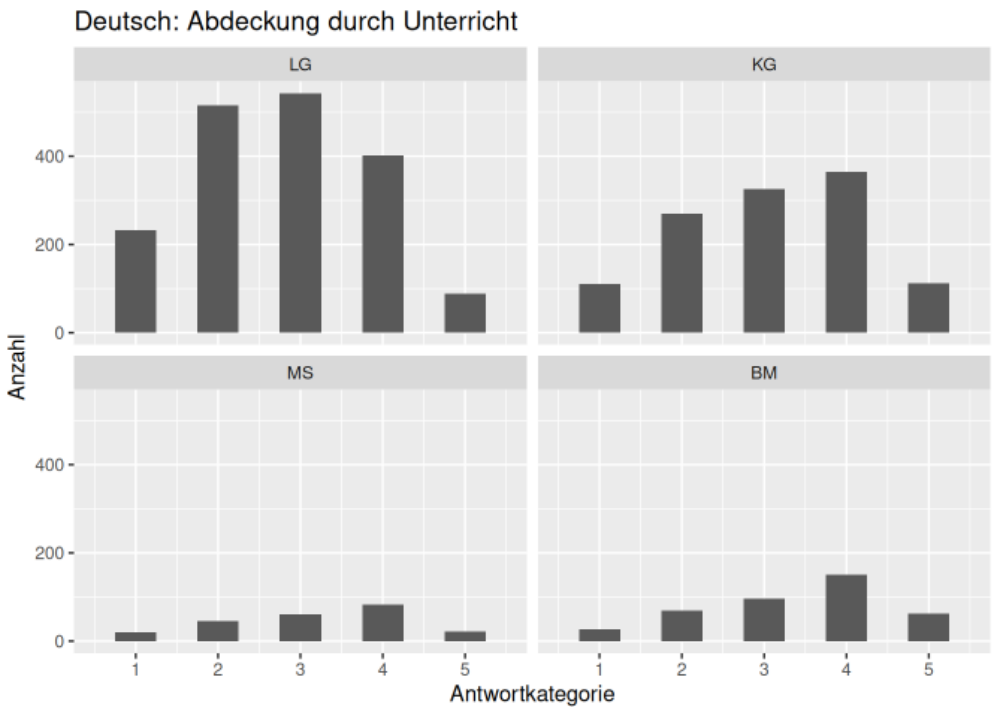


Abbildung 21

Abdeckung des Mathematik-Prüfungsstoffs durch einen Vorbereitungskurs nach den einzelnen Prüfungstypen und Antwortkategorien

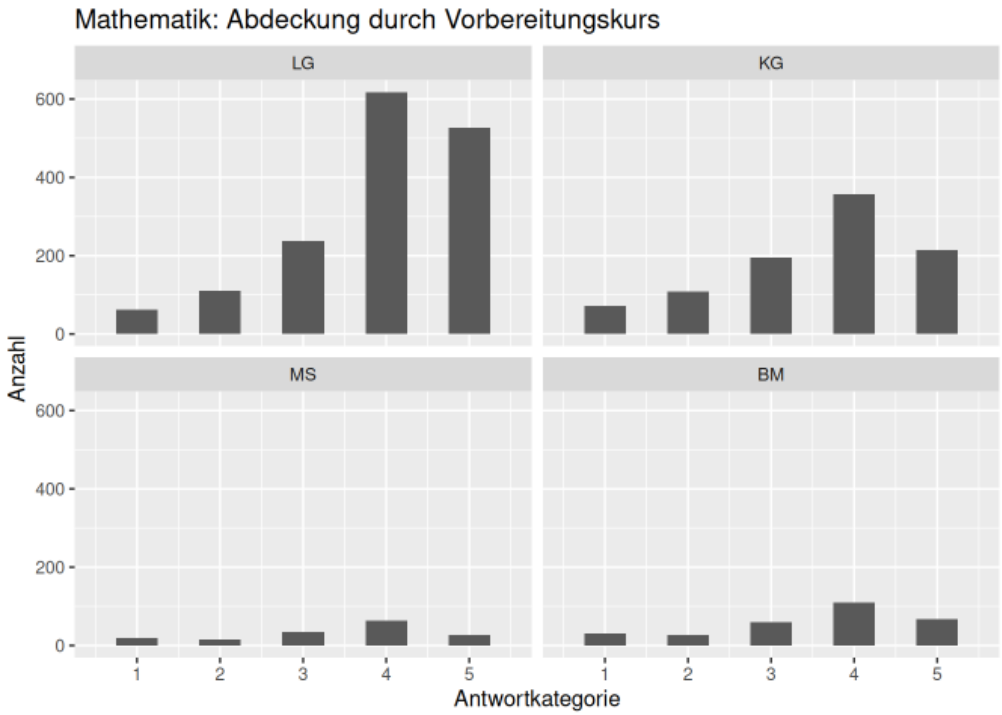
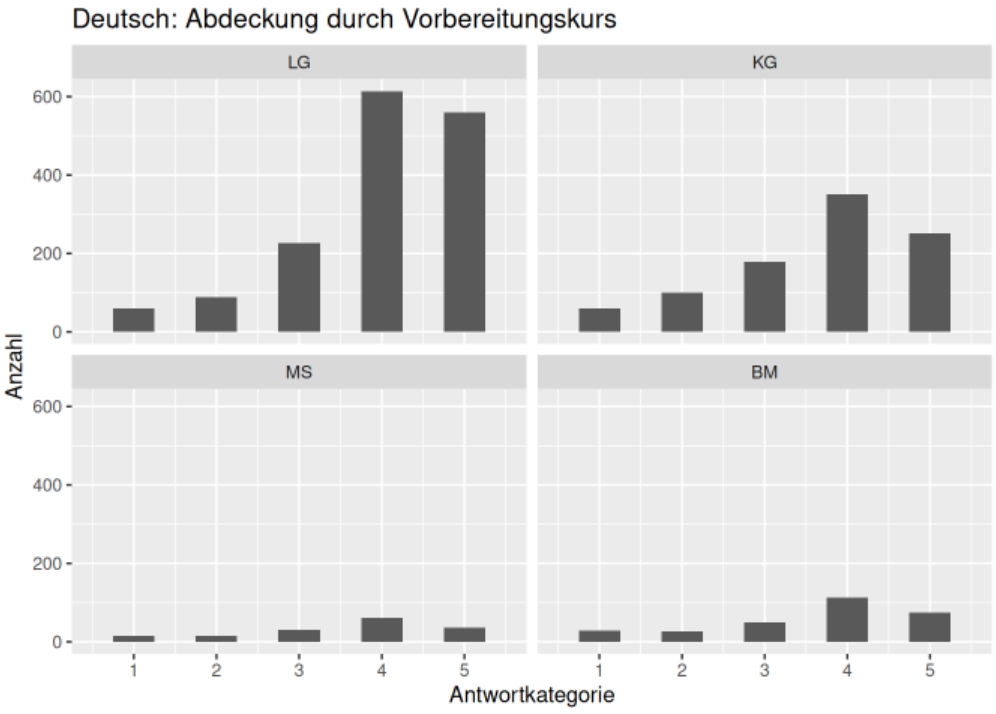


Abbildung 22

Abdeckung des Deutsch-Prüfungsstoffs durch einen Vorbereitungskurs nach den einzelnen Prüfungstypen und Antwortkategorien



Insgesamt fällt bei der Betrachtung der Unterschiede zwischen den Prüfungstypen auf, dass sich die Einschätzungen der Abdeckung des Prüfungsstoffs durch den Unterricht bzw. durch den Vorbereitungskurs gegenläufig verhalten. Für die LG-Prüfung wird die Stoffabdeckung durch den Unterricht im Vergleich zu den übrigen Prüfungen tendenziell geringer eingeschätzt, während die Abdeckung durch den Vorbereitungskurs hoch wahrgenommen wird. Diese Differenz kann inhaltliche Gründe haben (z. B. eine stärkere Passung der Kursinhalte zur Prüfung bzw. eine geringere explizite Prüfungsorientierung des Unterrichts). Sie könnte jedoch teilweise auch mit altersbezogenen Unterschieden in metakognitiven Fähigkeiten zusammenhängen: Da die Kandidierenden der LG-Prüfung im Durchschnitt jünger sind als jene der übrigen Prüfungen, ist plausibel, dass sie die Prüfungsrelevanz und den Transfer von Unterrichtsinhalten weniger sicher einschätzen und Unterricht daher eher als „weniger abdeckend“ erleben, während massgeschneiderte Formate wie Vorbereitungskurse als direkter und eindeutiger prüfungsbezogen wahrgenommen werden (Schneider, 2008; Weil et al., 2013).

7. Bestehen der Probezeit

Die bisherigen Ergebnisse betreffen Wahrnehmung und Vorbereitung vor der Prüfung. Abschliessend wird betrachtet, in welchem Zusammenhang diese und andere Merkmale mit dem späteren Bestehen der Probezeit stehen. Dafür sollen die Vornoten, die Prüfungsnoten, der Besuch eines öffentlichen oder privaten Vorbereitungskurses sowie einige soziodemographische Variablen als Prädiktoren verwendet werden. Ausserdem wird die Bestehenswahrscheinlichkeit auf den unterschiedlichen regionalen Ebenen untersucht. Diese Auswertungen beruhen ausschliesslich auf Daten der Prüfungsjahrgänge 2023 sowie 2024.[3] Für diesen Jahrgang lagen zum Zeitpunkt der Datenauswertung Angaben zum Bestehen der Probezeit aus 9'548 eindeutig als bestanden oder nicht bestanden auswertbaren Prüfungen zusammen mit den Vor- und Prüfungsnoten vor. Zudem hat es aus der Onlinebefragung des Prüfungsjahrgangs 2024 Angaben zum Besuch eines öffentlichen oder privaten Vorbereitungskurses für 1’756 Kinder und Jugendliche.

Aus dem Langgymnasium liegen post-VAM (2023-2024) insgesamt Daten von 4’549 Kindern vor. Die Bestehensquote der Probezeit liegt in dieser Population bei 94.9 Prozent. Für 1’023 Kinder liegen aus der Onlinebefragung des Prüfungsjahrgangs 2024 auch Angaben zum Besuch eines Vorbereitungskurses vor, wobei die Bestehensquote in dieser Teilstichprobe bei 94.6 Prozent liegt. Der geringe Unterschied in den Bestehensquoten zwischen Population und Stichprobe deutet darauf hin, dass die Ergebnisse aus der Onlinebefragung auch auf die Kinder generalisiert werden können, die nicht an der Onlinebefragung teilgenommen haben.

[3] Aufgrund unterschiedlicher Definitionen und Auswertungskriterien können die hier publizierten Ergebnisse von den kantonalen Probezeitstatistiken abweichen.

Für das Kurzgymnasium liegen post-VAM (2023-2024) insgesamt Daten von 3'526 Jugendlichen vor. Die Bestehensquote liegt bei 91.2 Prozent. Für 605 Jugendliche liegen auch Angaben zum Besuch eines Vorbereitungskurses vor, wobei die Bestehensquote dort bei 91.2 Prozent liegt. Auch hier gibt es praktisch keinen Unterschied in den Bestehensquoten.

Für die Mittelschulen wurden die Daten aus den Handelsmittelschulen, den Fachmittelschulen und den Informatikmittelschulen zusammengefasst, da die Stichproben für die Auswertungen sonst zu klein gewesen wären. Hier liegen post-VAM (2023-2024) insgesamt Daten von 1'285 Jugendlichen vor. Die Bestehensquote liegt bei 91.4 Prozent. Für 119 Jugendliche liegen auch Angaben zum Besuch eines Vorbereitungskurses vor, wobei die Bestehensquote dort bei 91.6 Prozent liegt. Der Unterschied in den Bestehensquoten ist auch hier gering.

Für die Berufsmaturität 1 liegen post-VAM (2023-2024) insgesamt Daten von 188 Jugendlichen vor. Die Bestehensquote liegt bei 89.9 Prozent. Für 9 Jugendliche liegen auch Angaben zum Besuch eines Vorbereitungskurses vor, wobei die Bestehensquote dort bei 100 Prozent liegt. Diese Teilstichprobe ist so klein, dass hier keine sinnvollen Auswertungen möglich sind.

7.1. Vorhersage durch Vornote und Prüfungsnote

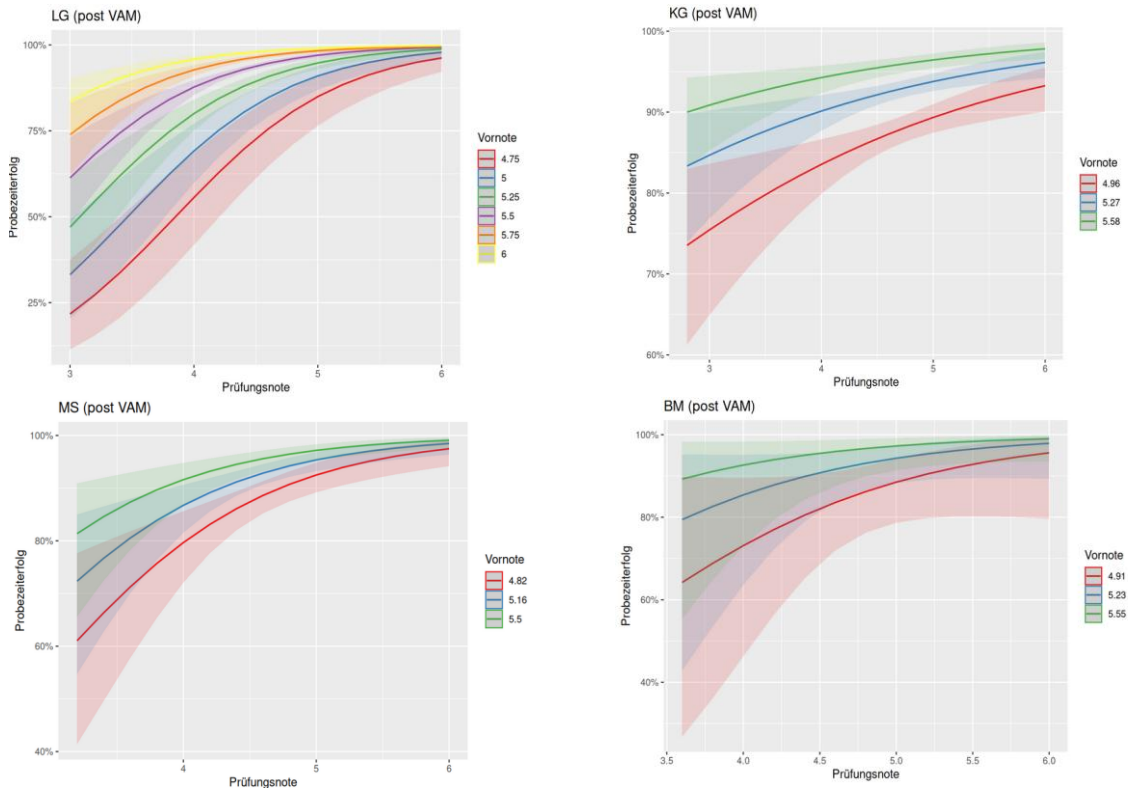
Sowohl die Vornote als auch die Prüfungsnote spielen bei der Vorhersage des Probezeiterfolgs eine Rolle, wenn auch in unterschiedlicher Stärke für die unterschiedlichen Prüfungstypen. Für das Langgymnasium beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.10. Die Odds Ratios betragen für die Vornote *OR* = 10.22 (p < .001) und für die Prüfungsnote *OR* = 4.51 (p < .001). Für das Kurzgymnasium beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.39. Die Odds Ratios betragen für die Vornote OR = 6.65 (p < .001) und für die Prüfungsnote *OR* = 1.65 (*p* < .001). Für die Mittelschulen beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.56. Die Odds Ratios betragen für die Vornote *OR* = 4.52 (*p* < .001) und für die Prüfungsnote *OR* = 3.15 (*p* < .001). Für die Berufsmaturität 1 beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.51. Die Odds Ratios betragen für die Vornote *OR* = 10.98 (*p* < .05) und für die Prüfungsnote *OR* = 2.83 (n.s.). Die Odds Ratios lassen sich in **Abbildung 23** so interpretieren, dass beispielsweise für das Langgymnasium eine um einen Notenpunkt höhere Vornote die Wahrscheinlichkeit des Bestehens der Probezeit etwa verzehnfacht.

Für die Mittelschulen, die Berufsmaturität 1 sowie mit Abstrichen auch für das Kurzgymnasium gilt, dass sich der Probezeiterfolg gut durch die zwei Noten vorhersagen lässt. Dabei spielt die Vornote in jedem Fall eine deutlich grössere Rolle als die Prüfungsnote. Für das Langgymnasium ist zwar das relative Gewicht der beiden Noten ähnlich wie bei den anderen Prüfungstypen. Dafür ist die Varianzaufklärung eher gering, was dafürspricht, dass die Noten insgesamt wenig aussagekräftig sind, um den Probezeiterfolg vorherzusagen.

Die Bedeutung der Vornote wird deutlich, wenn man sie als alleinigen Prädiktor für die Vorhersage des Probezeiterfolgs verwendet. Dabei unterscheidet sich die Varianzaufklärung nur geringfügig von der Varianzaufklärung in den Modellen, bei denen beide Noten berücksichtigt worden sind. Für das Langgymnasium sinkt die Varianzaufklärung von R^2 = 0.10 auf R^2 = 0.06 und das Odds Ratio für die Vornote liegt bei OR = 11.16 (p < .001). Für das Kurzgymnasium sinkt die Varianzaufklärung von R^2 = 0.39 auf R^2 = 0.39 und das Odds Ratio für die Vornote liegt bei *OR* = 6.85 (*p* < .001). Für die Mittelschulen sinkt die Varianzaufklärung von R^2 = 0.56 auf R^2 = 0.54 und das Odds Ration für die Vornote liegt bei *OR* = 4.44 (p < .001). Und für die Berufsmaturität sinkt die Varianzaufklärung von R^2 = 0.51 auf R^2 = 0.49 und das Odds Ration für die Vornote liegt bei *OR* = 12.75 (p < .01). Ginge es also lediglich um die Vorhersage des Probezeiterfolgs, hätte die Vornote allein fast die gleiche Vorhersagekraft wie die Vornote und die Prüfungsnote zusammen.

Abbildung 23

Odd-Rations des Probezeiterfolgs nach dem Regimewechsel nach den einzelnen Prüfungstypen, Prüfungsnoten und Vornoten



Umgekehrt scheint die Prüfungsnote selbst kaum Vorhersagekraft zu haben. Für das Langgymnasium sinkt die Varianzaufklärung von R^2 = 0.10 auf R^2 = 0.06 und das Odds Ratio für die Prüfungsnote liegt bei *OR* = 4.41 (*p* < .001). Für diesen Prüfungstyp scheinen beide Noten also eine ähnliche, aber nur sehr geringe Bedeutung zu haben. Für das Kurzgymnasium sinkt die Varianzaufklärung von R^2 = 0.39 auf R^2 = 0.02 und das Odds Ratio für die Prüfungsnote liegt bei *OR* = 1.83 (*p* < .001). Für die Mittelschulen sinkt die

Varianzaufklärung von R^2 = 0.56 auf R^2 = 0.04 und das Odds Ratio für die Prüfungsnote liegt bei *OR* = 2.67 (p < .001). Und für die Berufsmaturität sinkt die Varianzaufklärung von R^2 = 0.51 auf R^2 = 0.10 und das Odds Ratio für die Prüfungsnote liegt bei *OR* = 4.01 (p < .01).

Auch wenn die Prüfungsnote für den langfristigen Erfolg auf einer weiterführenden Schule durchaus diagnostisch sein kann (was in dieser Evaluation nicht überprüft worden ist, da es mit den vorhandenen Daten nicht überprüft werden kann), spielt sie für das Bestehen der Probezeit so gut wie keine Rolle. Indirekt rechtfertigt das die Institution einer Probezeit, da über die Prüfung und über die Probezeit offenbar unterschiedliche Fähigkeiten gemessen werden. Damit werden unterschiedliche Informationen genutzt, was zumindest im Prinzip zu einer besseren Vorhersage des langfristigen Erfolgs führen kann. Zu beachten ist ausserdem, dass diese Auswertungen nicht direkt darauf schliessen lassen, dass die Prüfung nicht nötig wäre. Die hier untersuchte Stichprobe zeichnet sich dadurch aus, dass sie über die Prüfung ausgewählt worden ist, überhaupt eine Probezeit machen zu können. Deswegen lassen sich die Befunde nicht einfach auf eine hypothetische Stichprobe übertragen, die prüfungsfrei die Probezeit absolviert hätte.

Versuchsweise wurden die Auswertungen auch mit sogenannten Regressionsbäumen durchgeführt. Hierbei wird für die einzelnen Prädiktorvariablen diejenige Ausprägung ausgewählt, die eine maximale Trennung der Gruppe der Erfolgreichen von der Gruppe der Nichterfolgreichen trennt. Der Vorteil dieser Auswertungen ist, dass man einen Punktwert erhält, bei dem diese Trennung jeweils maximal ist.

Für das Langgymnasium zeigt sich lediglich ein signifikanter Prädiktor. Bei einer Prüfungsnote unter 4.70 liegt die Wahrscheinlichkeit für das Bestehen der Probezeit bei 93 Prozent. Das betrifft 54 Prozent der Stichprobe. Bei einer Prüfungsnote von 4.70 und höher beträgt die Wahrscheinlichkeit für das Bestehen der Probezeit 98 Prozent. Eine solche Note haben 46 Prozent der Stichprobe. Insgesamt ist die Differenz in den Bestehensquoten aber nicht sehr gross, sodass sich auch hier wieder zeigt, dass sich das Bestehen der Probezeit auf dem Langgymnasium nur schlecht durch diese Prädiktoren vorhersagen lässt.

Für das Kurzgymnasium scheint es einen Interaktionseffekt zwischen Vornote und Prüfungsnote zu geben, sodass diese beiden Prädiktoren ausgewählt wurden. Bei einer Vornote unter 4.90 beträgt die Bestehenswahrscheinlichkeit 79 Prozent. Das betrifft 10 Prozent der Stichprobe. Bei einer Vornote von 4.90 und höher, welche 90 Prozent der Stichprobe erreicht haben, kommt es anschliessend auf die Prüfungsnote an. Ist diese unter 5.20 beträgt die Bestehenswahrscheinlichkeit für 71 Prozent der Stichprobe 90 Prozent. Ist diese jedoch bei 5.20 und höher, dann beträgt die Bestehenswahrscheinlichkeit 98 Prozent. Dies betrifft 20 Prozent der Stichprobe.

Für die Mittelschulen schliesslich spielt wieder lediglich die Vornote eine Rolle. Liegt diese unter 4.90, was bei etwa 15 Prozent der Stichprobe vorliegt, dann beträgt die Bestehenswahrscheinlichkeit lediglich etwa 80 Prozent. Liegt diese bei 4.90 oder höher,

was 85 Prozent der Stichprobe betrifft, dann beträgt die Bestehenswahrscheinlichkeit 92 Prozent.

Insgesamt sind die Befunde so zu lesen, dass Vornote und Prüfungsnote zwar beide mit dem Bestehen der Probezeit zusammenhängen, ihre diagnostische «Hebelwirkung» aber je nach Prüfungstyp unterschiedlich ausfällt. Besonders beim Langgymnasium ist zu berücksichtigen, dass die Bestehenswahrscheinlichkeit in der Probezeit insgesamt sehr hoch ist; in solchen Situationen sind Vorhersagemasse wie Pseudo-R^2 naturgemäss begrenzt, weil es schlicht wenig «Nichtbestehen» zu erklären gibt. Zudem wirkt hier eine Einschränkung der Varianz plausibel: Nach der Zulassungsentscheidung ist die Leistungsstreuung der tatsächlich eintretenden Kandidat*innen kleiner, wodurch Zusammenhänge im Nachhinein abgeschwächt erscheinen können (Hunter et al., 2006). Die Regressionsbäume liefern vor diesem Hintergrund keine «Grenzwerte» im Sinne von Entscheidregeln, sondern eine anschauliche, datengetriebene Beschreibung, in welchen Bereichen der Vornoten und Prüfungsnoten das Risiko des Nichtbestehens konzentrierter ist. Für die Praxis heisst das: Die Ergebnisse eignen sich gut bei Entscheidungen über Fördermassnahmen («Wo ballt sich Risiko?»), aber nicht zur direkten Ableitung neuer Schwellenwerte zu Selektionszwecken.

7.2. Verteilung der Bestehensquoten über Regionaleinheiten

Auf der tiefsten Regionalebene gibt es Daten von 656 Schulen. Die mittlere Wahrscheinlichkeit für das Bestehen der Prüfung pro Schule beträgt im Schnitt $M = 0.92$ ($SD = 0.14$), mit einer Spannweite von 0 bis 1, wie in **Abbildung 24** zu sehen ist. Auf der Gemeindeebene liegen uns 188 Schulgemeinden vor. Die durchschnittliche Bestehenswahrscheinlichkeit beträgt $M = 0.93$ ($SD = 0.11$), mit Werten zwischen 0 und 1. Auf Bezirksebene umfasst die Stichprobe 12 Bezirke. Die mittlere Bestehenswahrscheinlichkeit liegt bei $M = 0.93$ ($SD = 0.02$) mit einer Spannweite von 0.91 bis 0.97, wie in **Abbildung 25** ersichtlich ist.

Im Folgenden werden die Streuungen der Odds Ratios zwischen den unterschiedlichen Regionaleinheiten dargestellt. Auf der Schulebene waren diese Auswertungen nicht möglich, da manche Schulen so wenige Kandidierende abgegeben haben, dass die Parameterschätzungen instabil wurden. Dieses methodische Problem hätte dadurch gelöst werden können, dass man eine Mindestgruppegrösse definiert hätte. Diese Definition wäre jedoch immer willkürlich und die Ergebnisse dann auch nicht auf alle Schulen verallgemeinerbar. Aus diesem Grund wurde auf diese Auswertungen verzichtet.

Auf der Ebene von Schulgemeinden weist die Vornote ein mittleres Odd Ratio von $M = 24.31$ ($SD = 38.01$) mit einer Spanne zwischen 0.95 und 163.14 auf. Für die Prüfungsleistung ergibt sich ein mittleres Odds Ratio von $M = 2.97$ ($SD = 2.1$), mit einem Wertebereich von 0.78 bis 8.78. Das Streudiagramm zeigt, dass es einzelne Schulgemeinden gibt, für deren Schüler*innen die Vornote eine enorme Rolle für das Bestehen der Probezeit zu spielen scheint.

Auf der Ebene von Bezirken weist die Vornote ein mittleres Odd Ratio von *M* = 7.14 (*SD* = 6.50) mit einer Spanne zwischen 1.15 und 5.06 auf. Das Odds Ratio für die Prüfungsleistung hat auf dieser Ebene einen Durchschnitt von *M* = 2.58 (*SD* = 1.00), mit einem Wertebereich von 1.15 bis 5.06.

Abbildung 24

Dichte der mittleren Bestehenswahrscheinlichkeit auf Gemeindeebene

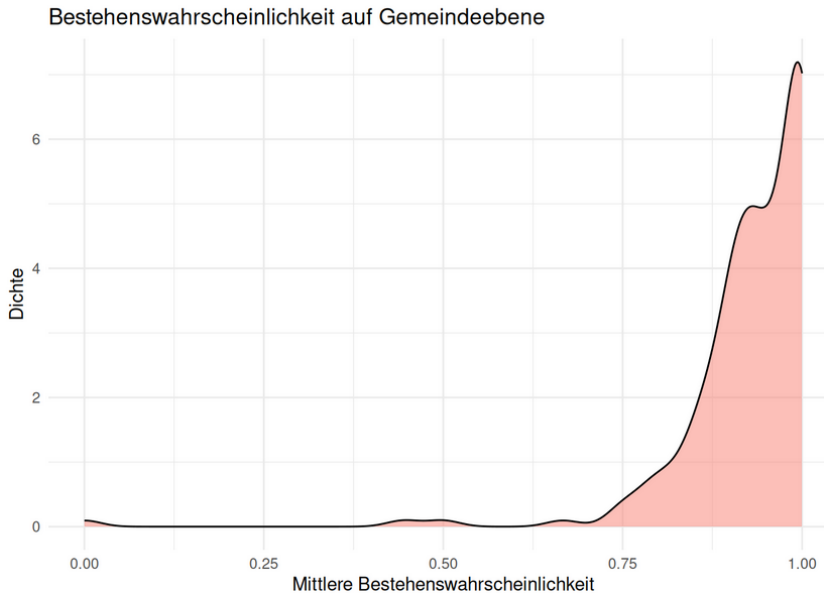
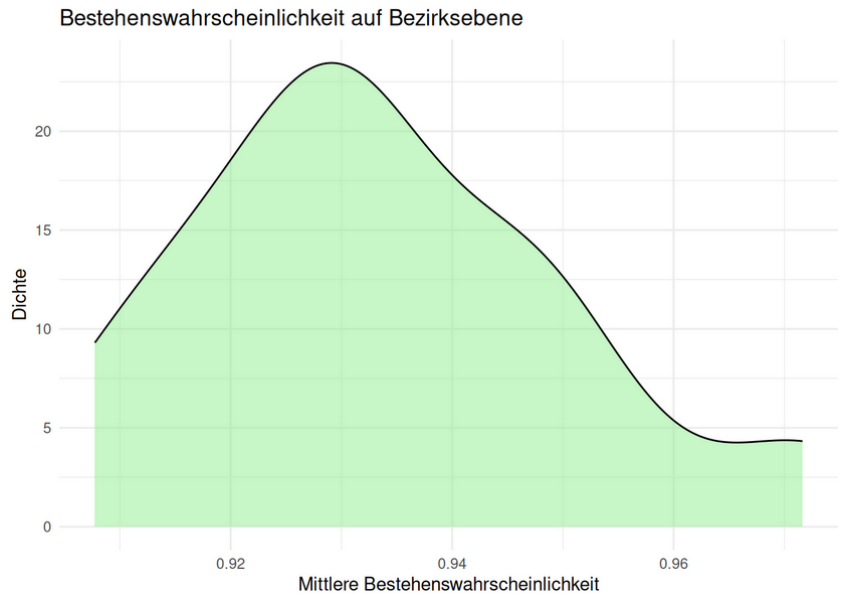


Abbildung 25

Dichte der mittleren Bestehenswahrscheinlichkeit auf Bezirksebene



Auf der Ebene von Bezirken liegt das Odds Ratio für Vornote im Mittel bei *M* = 7.14 (*SD* = 6.05), zwischen 1.47 und 24.19. Das Odds Ratio für die Prüfungsleistung hat hier einen Durchschnitt von *M* = 2.58 (*SD* = 1.00), mit einer Spanne von 1.15 bis 5.06. Im Streudiagramm sieht man keine grossen Auffälligkeiten, ausser dass es einen Schulbezirk zu geben scheint, bei dem die Vornote eine grosse Rolle spielt und die Prüfungsnote fast gar keine; und umgekehrt einen, bei dem die Prüfungsnote eine grosse Rolle spielt, die Vornote dagegen gar nicht.

Insgesamt zeigen die Befunde aus diesem Auswertungsschritt, dass sich die Effektstärken der Vornote und der Prüfungsnote zwischen den einzelnen Regionaleinheiten zuweilen deutlich unterscheiden. Das ändert jedoch nichts an dem Grundbefund, dass die Vornote in der Regel eine grössere Rolle spielt als die Prüfungsnoten. Wodurch sich die Streuung zwischen den Regionaleinheiten erklärt, müssen detaillierte Auswertungen zeigen.

Die Streuung der Bestehensquoten und der geschätzten Zusammenhänge über Schulen, Schulgemeinden und Schulbezirke zeigt vor allem eines: Die Vorhersage des Probezeit-Erfolgs durch Vornoten und Prüfungsnoten ist nicht überall gleich stark. Das kann verschiedene, rein deskriptiv zu verstehende Gründe haben – beispielsweise Unterschiede in Fallzahlen pro Einheit (was Quoten und Odds Ratios bei kleinen *N* stark volatil macht), Unterschiede in der Zusammensetzung der Kandidat*innen oder Unterschiede in schulischen Unterstützungs- und Übergangsbedingungen. Gerade weil in mehreren Einheiten sehr hohe Bestehenswahrscheinlichkeiten vorliegen, sollten grosse Odds Ratios nicht als «dramatisch hohe» Effektstärken gelesen werden, sondern als statistische Konsequenz geringer Ereignisraten bei gleichzeitig vorhandenen Leistungsunterschieden. Als datengeleitete Konsequenz ergibt sich daraus weniger eine kausale Interpretation, sondern ein Monitoring-Signal: Wenn sich die Heterogenität der Zusammenhänge über künftige Jahrgänge wiederholt, wäre eine vertiefte Modellierung mit kontextabhängigen Effekten (beispielsweise zufälligen Steigungen) angezeigt, um stabile Muster von Zufallsschwankungen zu trennen.

7.3. Prüfungsnoten vor und nach dem Regimewechsel

Im nächsten Schritt wurde überprüft, ob sich der Zusammenhang zwischen der Prüfungsnote und dem Bestehen der Probezeit durch den Regimewechsel verändert hat. Für das Langgymnasium beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.07 ante-VAM (2020-2022) und R^2 = 0.06 post-VAM (2023-2024). Das entsprechende Odds Ratio liegt bei *OR* = 5.22 ante-VAM (2020-2022) und bei *OR* = 4.41 post-VAM (2023). Für das Kurzgymnasium beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.21 ante-VAM (2020-2022) und R^2 = 0.02 post-VAM (2023-2024). Das entsprechende Odds Ratio liegt bei *OR* = 4.74 ante-VAM (2020-2022) und bei *OR* = 1.83 post-VAM (2023-2024). Für die Mittelschulen beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.13 ante-VAM (2020-2022) und R^2 = 0.04 post-VAM (2023-2024). Das entsprechende Odds Ratio liegt bei *OR* = 3.37

ante-VAM (2020-2022) und bei *OR* = 2.67 post-VAM (2023-2024). Für die Berufsmaturität 1 beträgt die Pseudovarianzaufklärung des Models nach Nagelkerke R^2 = 0.31 ante-VAM (2020-2022) und R^2 = 0.10 post-VAM (2023-2024). Das entsprechende Odds Ratio liegt bei *OR* = 17.87 ante-VAM (2020-2022) und bei *OR* = 4.01 post-VAM (2023-2024).

Während sich also für das Langgymnasium durch den Regimewechsel wenig geändert hat, zeigt sich für das Kurzgymnasium, die Mittelschulen und die Berufsmaturität 1, dass die Vorhersagekraft der Prüfungsnote für das Bestehen der Probezeit deutlich abgenommen hat. Dieser Befund ist erklärungsbedürftig. Zunächst ist zu berücksichtigen, dass sich die Rolle eines Prädiktors in einem multivariaten Modell verändern kann, sobald ein weiterer, mit ihm korrelierter Prädiktor berücksichtigt wird: Ein Teil der zuvor durch die Prüfungsnote erklärten Varianz kann dann durch die Vornote abgedeckt werden, sodass die Prüfungsnote einen kleineren einzigartigen Beitrag zur Vorhersage leistet (Hunsley & Meyer, 2003; Ludlow & Klein, 2014). Darüber hinaus ist denkbar, dass der Regimewechsel mit einer veränderten Zusammensetzung der Kandidat*innen einhergeht (z. B. andere Leistungsprofile unter den zugelassenen Kandidat*innen), was die Varianz der Prüfungsnote und damit ihre beobachtete Vorhersagekraft reduzieren kann (Hunter et al., 2006). Schliesslich ist der Befund auch mit einer veränderten Gewichtung von Leistungsindikatoren in der pädagogisch-administrativen Probezeitentscheidung vereinbar, wenn im neuen Regime zusätzliche Informationen (insbesondere Vornoten) systematisch stärker berücksichtigt werden.

In einem letzten Auswertungsschritt zur Vorhersage des Bestehens der Probezeit wurde der Besuch eines öffentlichen bzw. privaten Vorbereitungskurses mit dem Bestehen der Probezeit in Verbindung gebracht. Für die Mittelschulen sind die Ergebnisse sehr vorsichtig zu interpretieren. Dadurch dass hier die allermeisten Jugendlichen, die auch an der Onlinebefragung teilgenommen haben, auch die Probezeit bestanden haben, liegt eine quasi-perfekte Trennung vor und die geschätzten Modellparameter sind nicht stabil. Für die BM1 konnte kein Modell berechnet werden, da die Fallzahlen insgesamt viel zu klein waren. Ansonsten sind alle anderen Unterschiede nicht signifikant.

Mit einer Ausnahme zeigen sich in **Tabelle 10** keine signifikanten Zusammenhänge zwischen dem Besuch von Vorbereitungskursen und dem Bestehen der Probezeit. Die einzige Ausnahme stellt der Zusammenhang zwischen dem Besuch eines öffentlichen Vorbereitungskurses für das Kurzgymnasium dar. Dort verdoppelt der Besuch eines öffentlichen Vorbereitungskurses die Wahrscheinlichkeit, die Probezeit auch zu bestehen.

Die Ergebnisse verändern sich auch dann nicht wesentlich, wenn zusätzlich für Alter, Geschlecht, Erstsprache, Nationalität und Aufenthaltsdauer in der Schweiz kontrolliert wird. Der statistische Zusammenhang zwischen dem Besuch eines öffentlichen Vorbereitungskurses und dem Bestehen der Probezeit bleibt signifikant, während sich für die übrigen Variablen keine signifikanten Zusammenhänge zeigen. Eine kausale Interpretation ist jedoch nicht möglich, da die Zuweisung zum Besuch eines Vorbereitungskurses nicht randomisiert erfolgte. Es ist daher denkbar, dass sich

Schüler*innen mit unterschiedlichen Ausgangsvoraussetzungen systematisch in ihrer Kursnutzung unterscheiden. Beispielsweise könnten leistungsschwächere Schüler*innen häufiger einen Vorbereitungskurs besuchen, während dieser zugleich mit einer höheren Wahrscheinlichkeit des Bestehens der Probezeit einhergeht. Wenn sich solche gegenläufigen Zusammenhänge überlagern, kann sich im Gesamtbild kein klarer Zusammenhang zeigen.

Tabelle 10

Vorhersage des Bestehens der Probezeit durch den Besuch eines privaten vs. öffentlichen Vorbereitungskurses nach den einzelnen Prüfungstypen

Prüfung	Prädiktor	OR	CI-	CI+	p-Wert
LG					
LG	Privater Kurs	0.99	0.49	2.00	0.98
LG	Öffentlicher Kurs	0.88	0.38	2.00	0.75
KG					
KG	Privater Kurs	1.31	0.71	2.44	0.39
KG	Öffentlicher Kurs	2.29	1.24	4.24	0.01
MS					
MS	Privater Kurs	0.99	0.27	3.72	0.99
MS	Öffentlicher Kurs	2.05	0.49	8.52	0.32

Prüfung	Prädiktor	OR	CI-	CI+	p-Wert
LG					
LG	Privater Kurs	1.15	0.55	2.42	0.71
LG	Öffentlicher Kurs	0.96	0.41	2.25	0.93
KG					
KG	Privater Kurs	1.45	0.77	2.75	0.25
KG	Öffentlicher Kurs	2.28	1.20	4.33	0.01
MS					
MS	Privater Kurs	0.62	0.13	2.88	0.54
MS	Öffentlicher Kurs	2.86	0.56	14.62	0.21

Zusammenfassend ergibt sich ein konsistentes Bild mit zwei Ebenen: Erstens bleibt für das Langgymnasium die Rolle der Prüfungsnote für das Bestehen der Probezeit über den Regimewechsel hinweg weitgehend stabil, während sie in anderen Prüfungstypen post-VAM deutlich schwächer ausfällt. Dieser Befund ist inhaltlich anschlussfähig an zwei datennahe Lesarten: Entweder wird ein Teil dessen, was ante-VAM (2020-2022) über die Prüfungsnote «mit erklärt» wurde, post-VAM (2023-2024) bereits durch die (mit der Prüfungsnote korrelierte) Vornote abgedeckt (inkrementelle Validität; Hunsley & Meyer, 2003), oder die zugelassene Kandidat*innengruppe ist im neuen Regime in ihrer Prüfungsleistungsstreuung homogener, wodurch beobachtete Zusammenhänge kleiner werden (Hunter et al., 2006). Zweitens liefern die Analysen zu Vorbereitungskursen – bei aller Vorsicht wegen Modellinstabilität und fehlender Randomisierung – eher ein punktuelles als ein generelles Muster: Ein robuster Zusammenhang zeigt sich nur für den öffentlichen Vorbereitungskurs im Kurzgymnasium, während ansonsten keine konsistenten Unterschiede sichtbar werden. Insgesamt sind diese Befunde damit am besten als datengeleitete Orientierung zu lesen («Wo trägt welcher Indikator am meisten?»), nicht als abschliessende Aussage über Wirksamkeit; für belastbare Aussagen zur Stabilität post-VAM braucht es zwingend weitere Prüfungsjahrgänge.

8. Zusammenfassende Betrachtung

Zunächst sei festgehalten, dass die psychometrischen Skaleneigenschaften und die insgesamt gefundenen Korrelationsmuster der Skalen und Items untereinander für eine hohe Datenqualität sprechen. Die Jugendlichen haben für die Befragung zwischen 15 und 20 Minuten gebraucht und dabei offenbar sehr gewissenhaft gearbeitet, wie die Skalenanalysen gezeigt haben. Das erleichtert die Interpretation der vorliegenden Befunde. Damit ist auch die Grundlage gegeben, Unterschiede zwischen Prüfungstypen und Zusammenhänge zwischen Konstrukten als inhaltliche Muster zu lesen und weniger als Messartefakte, wobei einzelne Befunde aufgrund der teils kurzen Skalen weiterhin primär deskriptiv zu interpretieren sind.

Insgesamt bewerten die Jugendlichen, die an der Onlineumfrage teilgenommen haben, die Prüfung deutlich stärker als Herausforderung, denn als Bedrohung. Das korrespondiert mit einem durchaus positiven Erleben der Prüfungssituation, was sich in einem hohen positiven Affekt, einem niedrigen negativen Affekt, einer deutlich positiven Valenz und einer durchschnittlichen Erregung widerspiegelt. Im Durchschnitt ist die Prüfung den Jugendlichen wichtig und sie erleben auch relativ viel Kontrolle, die Prüfungssituation durch eigenes Zutun beeinflussen zu können. Dieses Zusammenspiel aus hoher Bedeutsamkeit und gleichzeitig eher herausfordernder (statt bedrohlicher) Bewertung ist in der Stress- und Bewältigungsforschung ein typisches Muster für Situationen, die als anspruchsvoll, aber grundsätzlich bewältigbar erlebt werden. Bei diesem Befundmuster sind die Unterschiede zwischen den Prüfungstypen meistens nicht sehr gross, stellenweise aber doch bemerkenswert. Dabei stechen einmal die LG-Prüfung und einmal die MS-Prüfung heraus. Jugendliche, die die LG-Prüfung absolviert haben, berichten von der geringsten Bedrohung und haben bei der emotionalen Bewertung die vorteilhafteste Einschätzung. Jugendliche, die die MS-Prüfung absolviert haben, berichten dagegen die höchste Bedrohung, was damit zusammenhängen mag, dass sie auch die Prüfung im Vergleich zu anderen Kandidat*innen als besonders wichtig wahrnehmen. Entsprechend fällt die emotionale Bewertung am wenigsten vorteilhaft aus. Passend dazu zeigte sich in den Detailanalysen, dass die subjektive Bestehenswahrscheinlichkeit eng mit erlebter Kontrolle und emotionalem Erleben zusammenhängt; der häufige Wert um 50 Prozent kann als Ausdruck von Unsicherheit direkt nach der Prüfung gelesen werden, trotz insgesamt eher positiver Grundstimmung.

Mehr als 90 Prozent der Jugendlichen gaben an, einen schulischen und/oder einen privaten Kurs zur Prüfungsvorbereitung zu nutzen, und fast jeder Fünfte nutzte sogar beide Angebote zusammen. Die Nutzung von Vorbereitungskursen hing dabei mit der kognitiven Bewertung der Prüfungssituation zusammen, wobei die Zusammenhänge komplex waren und sich aus ihnen keine kausalen Aussagen ableiten lassen. Dieser Punkt ist zentral: Kursnutzung ist sehr plausibel selbstselektiert, und die beobachteten Zusammenhänge sind deshalb primär als Muster der Nutzung und nicht als Wirksamkeitsnachweis zu interpretieren. Schulische Vorbereitungskurse werden besonders häufig von Jugendlichen besucht, die sich auf die LG-Prüfung vorbereiten, während sie besonders

selten von Jugendlichen genutzt werden, die sich auf die MS-Prüfung vorbereiten. Über mehrere Abschnitte hinweg ergibt sich damit ein konsistentes Bild: Dort, wo die Prüfung als besonders bedrohlich und bedeutsam erlebt wird (MS), ist die Nutzung privater Angebote besonders hoch; dort, wo das Erleben am vorteilhaftesten ist (LG), wird häufiger auf schulische Angebote zurückgegriffen.

Über die Nutzung von schulischen und privaten Vorbereitungskursen hinaus wurde ausserdem gefragt, ob sich die Jugendlichen durch ihre Schule, durch ihre Eltern und durch eine schulische Informationsveranstaltung für die Prüfung unterstützt gefühlt haben. Während die Unterstützung durch die Eltern als sehr hoch eingeschätzt wird, zeigt sich bei der Unterstützung durch die Schule und durch eine schulische Informationsveranstaltung ein gemischtes Bild. Bei den letzten Aspekten sind die Verteilungen ziemlich gleichverteilt, was darauf schliessen lässt, dass es Schulen gibt, die hier als sehr unterstützend wahrgenommen werden, und solche, bei denen diese Unterstützung gar nicht gesehen wird. Diese Polarität passt als Querverbindung zu den Befunden zur Kursnutzung: Wo schulische Unterstützung und Information als gering erlebt werden, ist es naheliegend, dass Jugendliche stärker auf externe/private Vorbereitung ausweichen; zugleich ist auch hier keine kausale Richtung ableitbar.

Schliesslich wurde gefragt, ob der Prüfungsstoff in der Einschätzung der Jugendlichen durch den regulären Unterricht bzw. durch einen Vorbereitungskurs abgedeckt gewesen ist. Bei beiden Aspekten zeigt sich ein gegenläufiger Befund. Die Abdeckung durch den Unterricht wird als gering angesehen, insbesondere für das Fach Mathematik. Dagegen wird die Abdeckung durch einen Vorbereitungskurs als sehr hoch angesehen. Dass die Zusammenhänge dieser Einschätzungen mit den tatsächlichen Prüfungsnoten eher klein ausfallen, ist damit vereinbar: Die Items erfassen primär wahrgenommene Passung und Prüfungsnähe (subjektive Abdeckung), nicht automatisch das Leistungsniveau.

Die Jugendlichen erachten die Prüfung also insgesamt als positiv und setzen bei der Vorbereitung auf schulische und/oder private Vorbereitungskurse, vermutlich weil sie dort eine höhere Abdeckung des Prüfungsstoffs erwarten als durch den regulären Unterricht und auch nicht immer ausreichend Unterstützung durch ihre Schule erfahren. Ein scheinbares Paradox («positives Erleben, aber sehr hohe Vorbereitung») löst sich damit auf: Hohe Bedeutsamkeit und herausfordernde Bewertung können gleichzeitig mit intensiver Vorbereitung auftreten; Vorbereitung ist dann nicht Ausdruck von Panik, sondern von Ernsthaftigkeit und dem Wunsch, Kontrolle herzustellen. Über mehrere Abschnitte hinweg wiederholt sich zudem ein Muster der Differenzierung nach Prüfungstypen: Die LG-Prüfung fällt häufig durch ein vorteilhafteres Erleben auf, während die MS-Prüfung in mehreren Indikatoren (Bedeutsamkeit/Bedrohung, Nutzung privater Kurse) als besonders belastet erscheint.

Schliesslich zeigte sich bei der Vorhersage des Bestehens der Probezeit durch Vornoten und Prüfungsnoten, dass es in erster Linie die Vornoten waren, die hier die grösste Vorhersagekraft hatten. Dieses Ergebnis ist vielleicht wenig überraschend, wenn man bedenkt, dass die Vornoten im Gegensatz zu den Prüfungsnoten näher an den

schulischen Anforderungen sind und somit neben allgemeinen kognitiven Fähigkeiten offenbar auch so etwas wie andauernde Anstrengungsbereitschaft, Sozialverhalten und andere Fähigkeiten messen, die im Schulalltag relevant sind. Im Zusammenspiel mit den Befunden aus Kapitel 1 (nur mässige Korrelation zwischen Vornote und Prüfungsnote) ergibt sich damit ein konsistentes Bild: Beide Grössen bilden unterschiedliche, komplementäre Aspekte von Leistung ab und für den längerfristigen Schulerfolg scheint der stärker alltagsnahe Anteil (Vornote) besonders bedeutsam.

Bei der Interpretation der Befunde ist zu beachten, dass hier lediglich Angaben von solchen Jugendlichen vorliegen, die an der Onlineumfrage teilgenommen haben. Mit einer hohen Wahrscheinlichkeit handelt es sich dabei nicht um eine Zufallsauswahl aller Jugendlichen, die an der Prüfung teilgenommen haben. Weitere Auswertungen mit Kenntnis der Vornoten und des Prüfungsergebnisses können dabei helfen, die wahrscheinliche Verzerrung in den Daten zu reduzieren. Damit ist die hier vorgelegte Zusammenfassung primär als Beschreibung der Muster innerhalb der befragten Teilgruppe zu verstehen; die Richtung der Befunde ist informativ, die genaue Höhe der Mittelwerte und Korrelationen kann jedoch durch Selektions- und Teilnahmeeffekte beeinflusst sein.

Literaturverzeichnis

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the sematic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*, 49–59. doi:10.1016/0005-7916(94)90063-9

Gaab, J. (2009). PASA – Primary Appraisal Secondary Appraisal: Ein Fragebogen zur Erfassung von situationsbezogenen kognitiven Bewertungen. *Verhaltenstherapie, 19*, 114–115. doi:10.1159/000223610

Heckhausen, J., Wrosch, C., & Schulz, R. (2010). A motivational theory of life-span development. *Psychological Review, 117*, 32–60. doi:10.1037/a0017668

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455. doi:10.1037/1040-3590.15.4.446

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594–612. doi:10.1037/0021-9010.91.3.594

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer Publishing.

Ludlow, L., & Klein, K. (2014). Suppressor variables: The difference between “is” versus “acting as”. *Journal of Statistics Education, 22*.

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education, 2*, 114–121. doi:10.1111/j.1751-228X.2008.00041.x

Tomasik, M. J., Silbereisen, R. K., & Pinquart, M. (2010). Individuals negotiating demands of social change: A control-theoretical approach. *European Psychologist, 15*, 246–259. doi:10.1027/1016-9040/a000064

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070. doi:10.1037/0022-3514.54.6.1063

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition, 22*, 264–271. doi:10.1016/j.concog.2013.01.004

Glossar

Martin J. Tomasik

Effektstärke

Die Effektstärke misst, wie stark der Unterschied oder Zusammenhang zwischen zwei Gruppen oder Variablen ist. Sie hilft dabei, die praktische Bedeutung eines Ergebnisses zu beurteilen, unabhängig von der Stichprobengrösse (vgl. Ellis, 2010). Cohens *d* ist eine der gängigsten Methoden zur Berechnung der Effektstärke (vgl. Cohen, 1988). Er gibt an, wie viele Standardabweichungen der Unterschied zwischen den Mittelwerten von zwei Gruppen beträgt. Üblicherweise redet man in den Sozial- und Verhaltenswissenschaften bei $d = .20$ von einem kleinen Effekt (d. h. ein schwacher, aber noch wahrnehmbarer Unterschied, bei $d = .50$ von einem mittleren Effekt (d. h. ein merklicher Unterschied, der in der Praxis oft relevant ist) und bei $d = .80$ von einem grossen Effekt (d. h. ein sehr deutlicher und praktisch fast immer relevanter Unterschied).

Einzelvergleich

Ein Einzelvergleich wird durchgeführt, nachdem eine Varianzanalyse (kurz ANOVA, von engl. «analysis of variance») mit mehr als zwei Gruppen zeigt, dass es signifikante Unterschiede zwischen den Gruppen gibt. Die Varianzanalyse selbst sagt nur, ob es irgendwo im Datensatz Unterschiede gibt, aber nicht, welche Gruppen sich genau unterscheiden. Die Einzelvergleiche (oft Post-hoc-Tests genannt) dienen dazu, diese Unterschiede zwischen den einzelnen Gruppen zu identifizieren und zu analysieren.

Haupteffekt

Ein Haupteffekt in einer ANOVA beschreibt den Einfluss einer einzelnen unabhängigen Variable auf die abhängige Variable, ohne die Interaktion mit anderen Variablen zu berücksichtigen. Er zeigt, ob es einen signifikanten Unterschied in den Mittelwerten der verschiedenen Gruppen gibt, die durch diese eine Variable definiert sind. Siehe auch Interaktionseffekt.

Interaktionseffekt

Ein Interaktionseffekt in einer ANOVA tritt auf, wenn der Effekt einer unabhängigen Variablen auf die abhängige Variable davon abhängt, wie eine andere unabhängige Variable ausgeprägt ist. Anders gesagt: Es wird überprüft, ob die Wirkung einer Variablen durch eine andere beeinflusst wird, sodass die Wirkung nicht einfach additiv ist. Siehe auch Haupteffekt.

Intraclasskorrelation

Mit der Intraclasskorrelation (ICC) lässt sich quantifizieren, wie ähnlich sich die Einheiten (z. B. Schüler) innerhalb einer bestimmten Gruppe oder Klasse sind. In Bezug auf Schulleistungen wird die ICC dazu verwendet, herauszufinden, wie viel der Leistung der Schüler durch die Schule beeinflusst wird, zu der sie gehören. Wenn die ICC hoch ist, bedeutet das, dass die Schulleistungen der Schüler innerhalb jeder Schule sehr ähnlich

sind und es nur geringe Unterschiede zwischen den Schulen gibt. Eine niedrige ICC hingegen zeigt an, dass die Leistungen der Schüler innerhalb einer Schule stark variieren und die Unterschiede zwischen den Schulen relativ klein sind. Mit anderen Worten: Eine hohe ICC würde darauf hinweisen, dass die Schule einen grossen Einfluss auf die Schulleistungen hat, während eine niedrige ICC darauf hindeutet, dass die Schülerleistungen mehr durch individuelle oder andere Faktoren beeinflusst werden als durch die Zugehörigkeit zu einer bestimmten Schule

Konfidenzintervall

Ein Konfidenzintervall ist ein Bereich, in dem ein wahrer Populationsparameter (z. B. Mittelwert oder Prozentsatz) mit einer bestimmten Wahrscheinlichkeit liegt. Es gibt also an, wie genau ein geschätzter Wert ist und bietet eine gewisse Unsicherheit oder Präzision der Schätzung. Ein Konfidenzintervall wird häufig mit einem Konfidenzniveau angegeben, das angibt, wie sicher man sich sein kann, dass der wahre Wert innerhalb des Intervalls liegt. Um Schätzer von zwei Gruppen miteinander zu vergleichen, kann man die Konfidenzintervalle der Mittelwerte oder anderer Parameter der beiden Gruppen betrachten. Wenn sich die Konfidenzintervalle der beiden Gruppen überlappen, deutet das darauf hin, dass es keinen signifikanten Unterschied zwischen den Gruppen gibt, weil der wahre Wert des Unterschieds zwischen den Gruppen mit hoher Wahrscheinlichkeit null sein könnte. Wenn sich jedoch die Konfidenzintervalle nicht überschneiden, spricht das für einen signifikanten Unterschied zwischen den Gruppen, da der wahre Unterschied mit hoher Wahrscheinlichkeit nicht null ist. Ein solches Ergebnis unterstützt die Hypothese, dass es tatsächlich einen Unterschied zwischen den beiden Gruppen gibt.

Korrelation

Korrelation beschreibt den statistischen Zusammenhang zwischen zwei Variablen – also, wie stark und in welche Richtung sie miteinander zusammenhängen. Eine Korrelation gibt an, ob und wie gut sich die Werte einer Variablen mit denen einer anderen Variablen verändern. Sie kann positiv, negativ oder null sein. Der Korrelationskoeffizient ist ein Wert, der den Grad und die Richtung dieser Beziehung quantifiziert. Der gängigste Korrelationskoeffizient ist der Pearson-Korrelationskoeffizient, der Werte im Bereich von -1 bis 1 annehmen kann. Ein (in der Praxis nie vorkommender) Korrelationskoeffizient von $r = 1.00$ bedeutet eine perfekte positive Korrelation. Wenn eine Variable steigt, steigt auch die andere Variable im gleichen Masse. Ein (in der Praxis auch nie vorkommender) Korrelationskoeffizient von $r = 1.00$ bedeutet eine perfekte negative Korrelation. Wenn eine Variable steigt, sinkt die andere im gleichen Masse. Ein Korrelationskoeffizient von $r = 0$ bedeutet, dass es zwischen zwei Variablen gar keinen Zusammenhang gibt. Es ist wichtig, zu beachten, dass Korrelation nicht Kausalität bedeutet. Nur weil zwei Variablen korreliert sind, heisst das nicht, dass die eine die andere verursacht. Es könnte auch ein dritter Faktor verantwortlich sein oder die Korrelation zufällig sein.

Mittelwert

Der Mittelwert, auch als Durchschnitt bezeichnet, ist eine zentrale Masszahl in der Statistik, die den «typischen» Wert einer Zahlenreihe darstellt. Er wird berechnet, indem man alle Werte einer Datenreihe addiert und die Summe durch die Anzahl der Werte teilt.

Signifikanz, statistische

Statistische Signifikanz bezieht sich darauf, ob ein beobachtetes Ergebnis in einer Studie wahrscheinlich auf einen tatsächlichen Effekt oder Unterschied zurückzuführen und nicht nur zufällig aufgetreten ist. In der Statistik wird ein Ergebnis als signifikant betrachtet, wenn die Wahrscheinlichkeit, dass es durch Zufall entstanden ist, unter einem vorher festgelegten Schwellenwert (dem Signifikanzniveau) liegt.

Signifikanzniveau

Das Signifikanzniveau oder auch der *p*-Wert ist der wichtigste Indikator für die statistische Signifikanz. Er gibt die Wahrscheinlichkeit an, dass das beobachtete Ergebnis (oder ein extremeres) unter der Annahme, dass keine wahre Wirkung existiert (d. h., die Nullhypothese zutrifft), zufällig auftritt. Ein *p*-Wert kleiner als .05 (häufig verwendete Schwelle, auch in diesem Bericht) bedeutet, dass das Ergebnis mit einer Wahrscheinlichkeit von weniger als 5 Prozent durch Zufall zustande gekommen ist und somit als statistisch signifikant gilt. Ein *p*-Wert grösser als .05 bedeutet, dass es keine ausreichenden Beweise gibt, um die Nullhypothese abzulehnen – das Ergebnis ist nicht signifikant.Standardabweichung

Die Standardabweichung ist ein Mass für die Streuung oder Variabilität der Werte in einer Datenreihe. Sie gibt an, wie stark die einzelnen Werte im Durchschnitt von ihrem Mittelwert abweichen. Eine geringe Standardabweichung bedeutet, dass die Werte nah beieinander und nahe dem Mittelwert liegen, während eine hohe Standardabweichung darauf hinweist, dass die Werte weit auseinander und stärker vom Mittelwert entfernt sind.

Varianzanalyse (ANOVA)

Die ANOVA ist eine statistische Methode, die verwendet wird, um Unterschiede zwischen den Mittelwerten von zwei oder mehr Gruppen zu testen. Ziel ist es herauszufinden, ob mindestens eine der Gruppen signifikant (siehe auch Signifikanz, statistische) von den anderen abweicht oder ob die beobachteten Unterschiede zufällig sind.

Universität Zürich[UZH]