Universität Zürich
Institut für Gymnasial- und Berufspädagogik

Franz Eberle und Mitarbeitende

Evaluation der Maturitätsreform 1995 (EVAMAR) Phase II

Kurzbericht zuhanden der EDK und des SBF

25. August 2008
(Stand vom 13.11.2008)

Inhaltsverzeichnis

1 Einleitung

Im Sommer 2001 beschlossen der Bund und die Konferenz der kantonalen Erziehungsdirektoren (EDK) eine gesamtschweizerische Evaluation (EVAMAR) der durch das MAR 95 (1995) eingeleiteten Reform der Maturitätsbildung. In der ersten Phase (EVAMAR I) wurden im Wesentlichen die folgenden drei Themen bearbeitet: (1) die Passung von Wahlfachangebot und Interessen der Schülerinnen und Schüler sowie die Auswirkungen auf den Ausbildungserfolg, insbesondere die subjektiv wahrgenommene Qualität der Vorbereitung auf ein Hochschulstudium; (2) die Umsetzung der fächerübergreifenden pädagogischen Ziele; (3) die Bewältigung der Reform durch die Schulen. Hauptinstrumente waren Befragungen, es erfolgten keine Leistungsmessungen. Die Ergebnisse von EVAMAR I liegen seit Ende 2004 vor. Sie bewerteten die Reform überwiegend als zufriedenstellend. Im Sommer 2005 beschlossen Bund und EDK den Beginn der zweiten Evaluationsphase. Im Projekt EVAMAR II sollte das Schwergewicht auf die objektivierte Erfassung des Ausbildungsstandes der Schülerinnen und Schüler am Ende des Gymnasiums gelegt werden.

Der Auftrag zur Projektleitung von EVAMAR II ging an Prof. Franz Eberle vom Institut für Gymnasial- und Berufspädagogik (IGB, vormals Höheres Lehramt Mittelschulen) der Universität Zürich. Dem Kernteam des IGB gehörten folgende wissenschaftliche Mitarbeitende an: Nina Barske, Karin Gehrer, Beat Jaggi, Johannes Kottonau, Maren Oepke und Michael Pflüger. Für die Romandie hat das IRDP Neuchâtel (Eva Roos) und für das Tessin das USR Bellinzona Aufgaben übernommen. An den Tests waren Urs Moser und Mitarbeitende vom Institut für Bildungsevaluation Zürich (IBE) massgeblich beteiligt, und im Zusammenhang mit dem überfachlichen Fähigkeitstest erfolgte eine Zusammenarbeit mit Prof. Klaus-Dieter Hänsgen vom Zentrum für Testentwicklung und Diagnostik Fribourg (ZTD). Das Teilprojekt „Analyse der Maturarbeiten – TP D2“ wurde vollständig durch die Pädagogische Hochschule der Fachhochschule Nordwestschweiz konzipiert und durchgeführt (Christina Huber, Prof. Vera Husfeldt, Lukas Lehmann und Prof. Carsten Quesel). Für spezifische Aufgaben war eine ganze Reihe von weiteren wissenschaftlichen Mitarbeitenden temporär ins Projekt eingebunden. Sie werden im Hauptbericht namentlich aufgeführt.

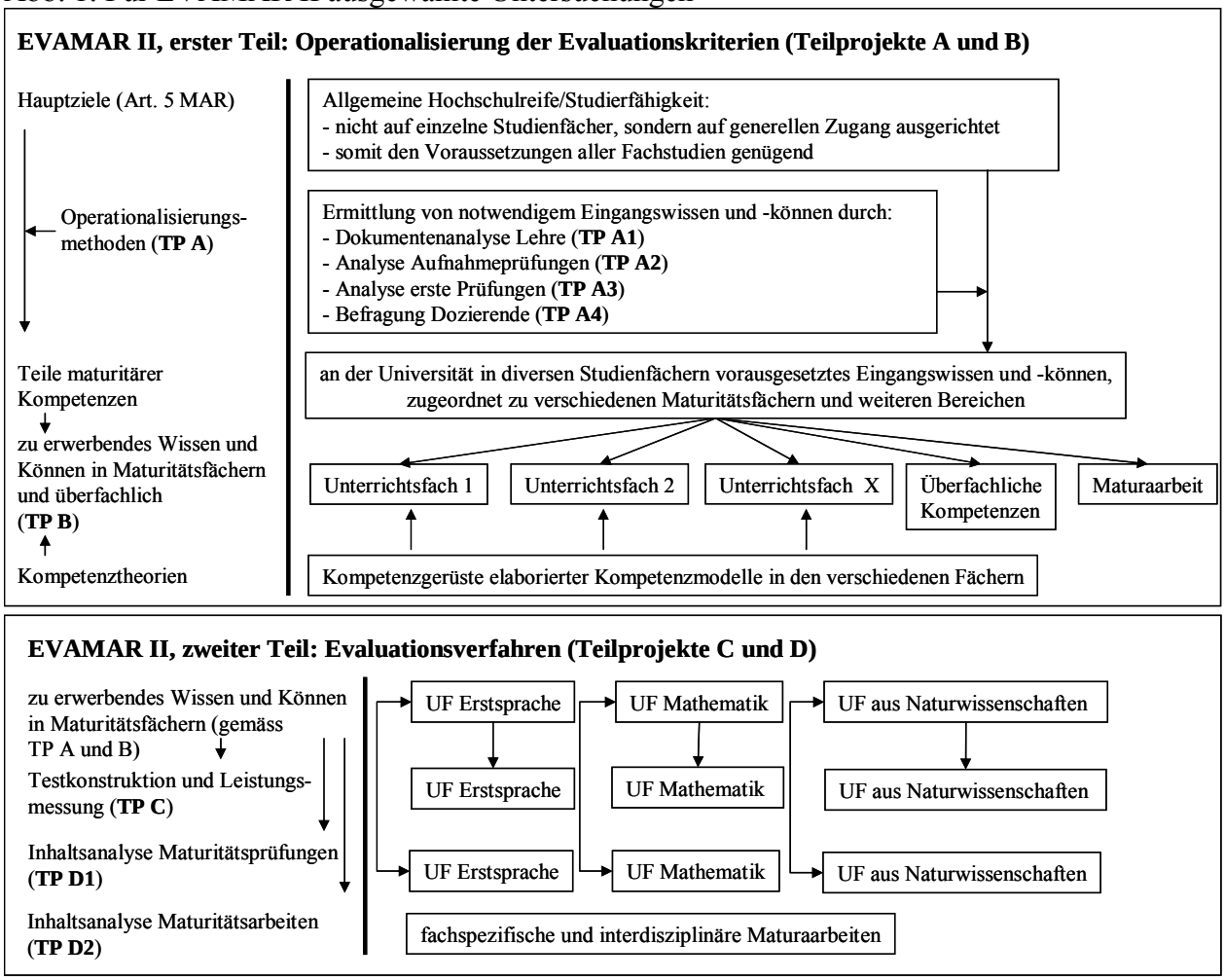
Der Projektauftrag beinhaltete die im nächsten Abschnitt beschriebenen Fragestellungen.

2 Fragestellungen

Eine evaluative Erfassung des Ausbildungsstandes muss sich sinnvollerweise an den Zielen der Ausbildung orientieren. Die Hauptziele der gymnasialen Bildung sind gemäss Art. 5 des MAR 95 (1995) die Erlangung „jener persönlichen Reife“, die erstens „Voraussetzung für ein Hochschulstudium ist“ (Hochschulreife oder Studierfähigkeit), und die zweitens „auf anspruchsvolle Aufgaben in der Gesellschaft vorbereitet“ (vertiefte Gesellschaftsreife durch breite Allgemeinbildung). Diese beiden Ziele haben teils gleichgerichtete, teils aber auch eigenständige und mitunter breite curriculare Auswirkungen. Eine vollständige Operationalisierung dieser Ziele in konkret zu erwerbende Kompetenzen der Gymnasiastinnen und Gymnasiasten in allen fachlichen und überfachlichen Lernbereichen und eine entsprechende flächendeckende Messung des Ausbildungsstandes, wie sie in einem ersten Konzept des Auftragnehmers vorgeschlagen wurde, wäre ausserordentlich aufwändig. Deshalb hat der Auftraggeber die im Rahmen von EVAMAR II vorzunehmenden Untersuchungen auf die in Abbildung 1 in ihrem Zusammenhang skizzierten Fragestellungen bzw. Teilprojekte eingeschränkt. Die Konkretisierung der Ziele der Maturitätsbildung sollte für ausgewählte Elemente des Hauptzieles der Studierfähigkeit erfolgen. Das zweite Hauptziel (vertiefte Gesellschaftsreife durch breite Allgemeinbildung) wurde damit nur teilweise in die durch den Auftraggeber gewählte Ziel-Operationalisierung einbezogen (im Überschneidungsbereich).

Auf der Grundlage der für EVAMAR II ausgewählten Untersuchungen können somit nur über *Teil*aspekte der Studierfähigkeit Aussagen gemacht werden. Das Ziel, „auf anspruchsvolle Aufgaben in der Gesellschaft vorzubereiten", wurde zudem nicht direkt untersucht (siehe auch die im Abschnitt 7 beschriebenen Grenzen der Untersuchung). Diese Einschränkungen sind bei der Interpretation der Ergebnisse und bei der Ableitung von Massnahmen für die künftige Gestaltung des Gymnasiums zu beachten!

Abb. 1: Für EVAMAR II ausgewählte Untersuchungen



Legende: - Die Buchstaben A-D entsprechen den Teilprojekten.
- UF: Abkürzung für „Unterrichtsfach"
- TP: Abkürzung für „Teilprojekt"

Die nachfolgende Darstellung der Ergebnisse fällt für die Teilprojekte A und B (Voruntersuchungen) knapper aus als für die Teilprojekte C und D, weil in dieser Kurzfassung des Berichts das Schwergewicht auf der Darlegung der Hauptergebnisse liegt.

3 Ergebnisse I: Voruntersuchung zur Konkretisierung des zu Messenden

3.1 Vorausgesetztes Wissen und Können zu Beginn eines Studiums

3.1.1 Inhaltsanalyse von Studienmaterialien und Prüfungen – TP A1 und A3

In den Teilprojekten A1 und A3 wurde eine Inhaltsanalyse von Lehrmaterialien und Prüfungen durchgeführt, um in den Lehrveranstaltungen des ersten und zweiten Semesters vorausgesetztes Wissen und Können zu ermitteln. Die Untersuchung erfasste die 16 im Studienjahr 2004/2005 meist gewählten Fächer, die von etwa 70% aller Immatrikulierten in

der Schweiz studiert wurden. Es wurden für jedes Fach jeweils jene zwei oder drei Universitäten angefragt, an denen zusammen mehr als 50% aller Studierenden dieses Faches eingeschrieben waren. Obwohl nicht alle kontaktierten Dozierenden ihre Studienunterlagen und Prüfungen bereit stellten, umfassten die eingegangen Unterlagen alle ausgewählten 16 Fächer. In Bezug auf die Analyse der ersten Prüfungen im Rahmen von TP A3 ist einschränkend noch festzuhalten, dass nicht für jedes der Studienfächer Prüfungen zur Verfügung standen, weil in manchen Studiengängen aufgrund der Studienstruktur innerhalb der analysierten ersten beiden Semester keine Prüfungen stattfanden.

Die Analyse von Unterrichtsmaterialien und Prüfungen zielte auf die in Abbildung 2 dargestellten Zusammenhänge zwischen Maturitäts- und Studienfächern. Die resultierende, in eine Datenbank eingelesene Sammlung von Wissens- und Könnenselementen bildet somit nicht einfach den Rahmenlehrplan und die Fachlehrpläne an Maturitätsschulen ab. Die in EVAMAR II angewandte Analyse- und Codiermethodik im Teilprojekt A (und B) musste neu entwickelt werden (siehe die ausführliche Beschreibung im Hauptbericht).

Abb. 2: Zusammenhang zwischen Maturitäts- und Studienfächern

Wissen und Können aus: →	Mathematik	Erstsprache	Geographie	Physik	Biologie	usw.: Bereich Y
für Studienfach ↓						
Germanistik	GM	GE	GG	GP	GB	GY
Recht	JM	JE	JG	JP	JB	JY
Soziologie	SM	SE	SG	SP	SB	SY
Biologie	BM	BE	BG	BP	BB	BY
usw.: X	XM	XE	XG	XP	XB	XY
	↓	↓	↓	↓	↓	↓
auszubildende Studierkompetenzen in	**Mathematik**	**Erstsprache**	Geographie	Physik	**Biologie**	Fach Y oder überfachl.
	↑	↑			↑	
elaborierte Kompetenzmodelle aus Mathematik, Erstsprache und Biologie						

Legende: - GM = Erforderliches Wissen und Können aus dem Maturitätsfach Mathematik (M) für das Studienfach Germanistik (G) usw.

Insgesamt kann man aus den untersuchten Lehrtexten und Prüfungen über alle Studienfächer hinweg auf folgende Tendenzen hinsichtlich des vorausgesetzten und anzuwendenden Wissens und Könnens aus den verschiedenen Maturitätsfächern schliessen:

- Bei der Zuordnung zu verschiedenen Wissenskategorien überwiegen die Einträge zur Kategorie Faktenwissen deutlich, vorausgesetztes prozedurales oder metakognitives Wissen wurde dagegen eher selten gefunden. Diese Häufung gilt insbesondere für Eintragungen bei den sprachlichen Schulfächern und betrifft daher zunächst die drei Erstsprachen Deutsch, Französisch und Italienisch sowie auch Englisch. Viele Eintragungen sind aber auch in den Fächern Latein und Griechisch zu finden, wobei es sich zum grossen Teil um Doppeleintragungen zusätzlich zur Erstsprache handelt. Darin widerspiegelt sich der Kontext der alten Sprachen zur Begrifflichkeit in vielen Fachwissenschaften.
- Die Zuordnung der Einträge zu den einzelnen Schulfächern fällt sehr unterschiedlich aus und hängt erwartungsgemäss stark vom analysierten Studienfach ab. So häufen sich z. B. bei der Analyse des Studienfachs Biologie Einträge bei den Schulfächern Biologie und Chemie, oder die wirtschaftswissenschaftlichen Studiengänge bauen stark auf ein bereits überraschend umfangreich vorausgesetztes, wirtschaftliches Begriffsrepertoire auf.

- Griechisch- und Lateinkenntnisse können überwiegend als hilfreich, aber nicht unbedingt als notwendig angesehen werden, so etwa in den Studiengängen Rechtswissenschaften, Geschichte oder Pädagogik; eine profunde Kenntnis von Latein und Griechisch dürfte jedoch vor allem – wie bereits erwähnt – Vorteile für den Wortschatz insgesamt mit sich bringen, der sich wiederum über alle Studiengänge hinweg als sehr bedeutsam erweist. Dabei ist aber zu berücksichtigen, dass Fachtermini auch ohne ein entsprechendes Sprachenlernen verständlich sein bzw. gemacht werden können.
- Neben den wichtigen Erstsprachkenntnissen und einem umfangreichen Fremdwortschatz wird von den Studierenden zum Teil ein gutes Leseverständnis gefordert: Insbesondere in den geistes- und sozialwissenschaftlichen Fächern Geschichte, Soziologie und Pädagogik häufen sich Eintragungen, die auf das Arbeiten mit komplexen Texten verweisen und eine entsprechende Lesekompetenz erfordern.
- Hinsichtlich der naturwissenschaftlichen Kenntnisse lassen sich bei allen nicht naturwissenschaftlichen Studienfächern (in unserem Falle alle Studienfächer ausser Physik, Biologie und Medizin) kaum Unterschiede in der Zahl der ermittelten Wissenselemente aus den Bereichen der Physik und der Chemie feststellen. Hingegen ist der Rückgriff auf biologische Begriffe und Konzepte in den Geistes- und Sozialwissenschaften deutlich häufiger als auf solche aus Physik und Chemie.
- Auch aus dem überfachlichen Bereich des Wissenschaftlichen Arbeitens werden Kenntnisse vorausgesetzt. Daneben werden aber auch in einigen Studienfächern Veranstaltungen angeboten, die in das Wissenschaftliche Arbeiten einführen.
- Die Analyse der Aufgaben und Prüfungen in Bezug auf das vorausgesetzte Wissen zeigt, dass meist auf wenig Vorwissen referiert wird, das bereits vom Gymnasium mitgebracht werden muss; als für die Prüfungen erforderliche kognitive Fähigkeiten wurden überwiegend die Fähigkeitsstufen „Erinnern von Wissen“, „Verstehen“, „Anwenden“ und „Analysieren“ befunden. Fähigkeiten des „Evaluierens, Kreierens“ oder des „Generierens und Planens“ werden hingegen von den Studierenden eher wenig gefordert, zumindest in den ersten beiden Studiensemestern.

Die in den Teilprojekten A1 und A3 gewonnenen Inhalte sollten vor allem als wichtigste Grundlage für die im weiteren Projektverlauf zu testenden Schulfächer Erstsprache, Mathematik sowie Biologie dienen.

3.1.2 Befragung von Dozierenden – TP A4

Ziel der Befragung von Dozierenden war es zu erfahren, welches fachliche und überfachliche Wissen und Können sie bei Studierenden des ersten und zweiten Semesters für ihre Veranstaltungen jeweils voraussetzen. Darüber hinaus wurden sie auch nach auffälligen Lücken im fachlichen und überfachlichen Wissen der Studienanfängerinnen und -anfänger gefragt. Angeschrieben wurden 310 Dozierende von Deutschschweizer Universitäten, deren Lehrveranstaltungen zuvor auch für die Analyse der Studienunterlagen ausgewählt worden waren. Die Rücklaufquote lag bei 45%. Die Hauptergebnisse, die von einigen Professoren der Romandie und des Tessins im Wesentlichen auch als für ihr Sprachgebiet zutreffend eingeschätzt wurden, sind nachfolgend aufgeführt.

Die Dozierenden wurden zunächst gebeten einzuschätzen, für wie wichtig sie fachspezifisches Wissen und Können aus den verschiedenen gymnasialen Maturafächern sowie in den Bereichen „Benutzerwissen Informatik“ und „Informatikfachwissen“ für das Verständnis und den erfolgreichen Besuch ihrer Veranstaltungen erachten. Eine Rangierung der Mittelwerte, gebildet aus einer vierstufigen Skala für die Antwortmöglichkeiten, ergibt die grösste Bedeutungszumessung für Deutsch, Englisch, Informatik-Benutzerwissen und Mathematik (in dieser Reihenfolge). Biologie und Physik befinden sich in dieser Reihung in der Mitte der

Rangskala und Chemie am Beginn des letzten Drittels. Die Prozentzahlen für die Antwortmöglichkeiten sind in Tabelle 1 ersichtlich.

Tab. 1: Bedeutung gymnasialer Fachbereiche für die eigene Lehrveranstaltung

gymnasiale Fächer oder Bereiche	kein Wissen und Können	fragmentarisches Wissen und Können	Kenntnis wichtigster Grundlagen	fundiertes Wissen und Können
(1) Deutsch (Erstsprache)	5.2%	6.0%	25.4%	63.4%
(2) Englisch	16.7%	11.9%	43.7%	27.8%
(3) Informatik-Benutzerwissen	19.7%	18.9%	45.9%	15.6%
(4) Mathematik	36.1%	15.6%	32.8%	15.6%
(10) Biologie	68.1%	11.8%	13.4%	6.7%
(11) Physik	70.1%	10.3%	14.5%	5.1%
(16) Chemie	78.8%	12.7%	2.5%	5.9%

Schlüsselt man die Antworten auf diese Frage nach den einzelnen Studienfächern auf, verbleibt Deutsch bzw. Erstsprache fast durchwegs an erster Stelle und Englisch sowie Informatik-Benutzerwissen auf den vorderen Rängen. Bei den anderen Fächern ergeben sich teilweise erhebliche Unterschiede in der Rangierung. Mathematik wird in den sprach- und geisteswissenschaftlichen Fächern erwartungsgemäss als wesentlich weniger wichtig eingestuft, bleibt aber für viele Studienfächer sehr bedeutsam. An vorderer Stelle ist jeweils jenes Maturafach zu finden, das dem eigenen Studienfach entspricht.

Bei einer weiteren Frage mussten die Dozierenden konkretes vorausgesetztes Wissen und Können aus den verschiedenen Fächern aufführen. Hier zeigte sich, dass Abweichungen zum curricularen Verständnis des jeweiligen Faches am Gymnasium bestehen können. Wenn Dozierende von der grossen Bedeutung von Deutsch sprechen, meinen sie vorwiegend „Sprachbeherrschung", Grammatik, Syntax u. ä.. Eigentliche Kerninhalte der Germanistik wie Sprach- oder Literaturgeschichte, sprachwissenschaftliche Interpretations- und Textanalysemethoden, Grundlagen der Linguistik, Dramatheorie usw. werden nur von einigen der Germanisten und einem Anglisten erwartet. Es verbergen sich also hinter dem Begriff „Fach Deutsch" unterschiedliche Vorstellungen hinsichtlich der Fachinhalte. Wenn Deutsch allgemein als wichtigstes Maturafach eingeschätzt wird, ist somit hauptsächlich deutsche Sprachkompetenz gemeint. Ähnlich verhält es sich beim Fach Englisch, bei dem die Fähigkeit zur Bearbeitung englischsprachiger Fachliteratur (und nicht etwa literaturgeschichtliche Kompetenz) im Mittelpunkt steht.

Die nächste Frage war jene nach Lücken im zuvor als wichtig bezeichneten Wissen und Können bei Studienanfängerinnen und -anfängern. In fast allen Fächern wurden solche genannt, herausragend viele in Erstsprache und Mathematik. In Erstsprache sind es dabei wieder im Wesentlichen Komponenten der „Sprachbeherrschung", deren „Dürftigkeit" vor allem Nicht-Germanistik-Dozierende monierten.

Die Dozierenden sollten in einer weiteren Frage auf einer siebenstufigen Skala einschätzen, für wie wichtig sie überfachliche Kompetenzen für das Verständnis und den erfolgreichen Besuch ihrer Veranstaltungen erachten. Alle vorgelegten Kompetenzen wurden als sehr wichtig (nahe bei der Stufe 7) bis mindestens „mittelwichtig" (leicht über Stufe 4) bewertet, und zwar in folgender Reihenfolge: (1) Selbstständiges Lernen, (2) Verantwortung für eigenes Lernen und Arbeiten übernehmen, (3) Selbstständiges Arbeiten, (4) Kritisches Denken, (5) Problemlösefähigkeit, (6) Schriftliche Ausdrucksfähigkeit, (7) Zeit effizient einteilen und Prioritäten setzen, (8) Mit Belastungen umgehen, (9) Hörverstehen im Zusammenhang mit Vorlesungen, (10) Mündliche Ausdrucksfähigkeit, (11) Umfangreiche Prüfungen vorbereiten,

(12) Entwickeln neuer Ideen, (13) In kurzer Zeit viel Lernstoff verarbeiten, (14) Recherchefähigkeit, (15) Im Team arbeiten, (16) Vor Publikum auftreten. Im Weiteren konnten die befragten Dozierenden angeben, bei welchen dieser überfachlichen Kompetenzen sie Defizite feststellten. Am häufigsten genannt wurden die Folgenden: Schriftliche Ausdrucksfähigkeit (42.4%), Kritisches Denken (35%), Selbstständiges Arbeiten (30%), Selbstständiges Lernen (26%), Mündliche Ausdrucksfähigkeit (21%), Verantwortung für eigenes Lernen und Arbeiten übernehmen (20.6%). Die wenigsten Nennungen erhielten: Vor Publikum auftreten (3%) und Im Team arbeiten (3%).

Die Befragungsergebnisse haben im Hinblick auf die Tests fast ausnahmslos keine Korrektur der Erkenntnisse aus den Analysen A1-A3 erforderlich gemacht, sind aber als eigenständige Ergebnisse des Hauptberichts interessant. Als Ausnahmen dazu sind festzuhalten: In Erstsprache wurde auch ein Bereich „Grammatikalische Kompetenzen/Orthografie“ in den Test für Erstsprache übernommen. In einem ergänzenden Fragebogen wurde um eine Selbsteinschätzung wichtiger überfachlicher Kompetenzen gebeten. Und schliesslich stützen die Befragungsergebnisse die Wahl von Biologie als naturwissenschaftliches Fach (siehe weiter hinten).

3.1.3 Analyse bestehender Aufnahmeprüfungen – TP A2

Mit diesem Teilprojekt sollten anhand einer Analyse von Eingangsprüfungen an Schweizer Universitäten ergänzende Hinweise zu den Kompetenzen ermittelt werden, die für ein universitäres Studium vorausgesetzt werden. Zu diesen Prüfungen gehören zunächst die Aufnahmeprüfungen, die an den meisten Universitäten als Qualifikationsmöglichkeit für den Studienzugang angeboten werden, wenn der Nachweis des Bestehens einer Schweizer Matura fehlt. Eine erste Übersicht hat ergeben, dass diese Prüfungen in der Regel an das Prüfungsprogramm der Schweizerischen Maturitätsprüfungen gekoppelt, also den Maturaprüfungen an den Gymnasien sehr ähnlich sind. Weil letztere aber Untersuchungsgegenstand des Teilprojekts D1 sind und deshalb die Gefahr von zirkulären Schlüssen bestanden hätte, musste diese Kategorie von Zugangsprüfungen von der weiteren Analyse ausgeschlossen werden. Als weitere Kategorie von Eingangsprüfungen bieten sich eigentliche Eignungstests an, die sich auf explizite, theoretisch begründete und möglichst auch empirisch validierte Studienerfolgsprognosekonzepte stützen. Es hat sich gezeigt, dass in der Schweiz der Eignungstest für das Medizinstudium (EMS) das einzige Instrument ist, das diesen Anforderungen genügt. Er wird für die auf die beschränkten Studienplätze ausgerichtete Zugangsselektion für das Medizinstudium an verschiedenen Universitäten eingesetzt und enthält Aufgaben, die überfachliche Kompetenzen erfassen (kein Wissenstest). Da der Einsatz solcher Aufgaben die Möglichkeit bot, in Ergänzung zu den in EVAMAR II vorgesehenen Fachtests weitere studienrelevante Kompetenzen zu erheben, entschied die Steuergruppe des Projekts EVAMAR II, im Rahmen des Teilprojekts C in der Deutschschweiz auch Aufgaben aus früheren Versionen des EMS vorzulegen. Dieser Testteil war im Auftragskonzept noch nicht vorgesehen.

3.2 Kompetenzmodellierung – TP B

In der Planung von EVAMAR II war zunächst vorgesehen, die Ergebnisse der Analysen der Teilprojekte A1-A4 direkt als Kompetenzmodelle zu verdichten. Dieses Vorhaben erwies sich aber schon bald aus verschiedenen Gründen als zu schwierig. Hauptsächlich deshalb, weil die in Projektteil A untersuchten Lehrtexte und Fachaufgaben zwar direkte Schlüsse auf jenes unabdingbare Wissen und Können zulassen, welches Voraussetzung für deren problemlose Bearbeitung ist, diese Informationen aber nicht ausreichend sind als Grundlagenelemente eines kohärenten, den aktuellen lerntheoretischen Erkenntnissen entsprechenden und empirisch abgesicherten Kompetenzmodells, wie es in der aktuellen theoretischen Diskussion

um Bildungsstandards vertreten wird. Zudem decken die ermittelten Sinn- oder Wissenseinheiten und kognitiven Fähigkeiten die Anforderungen der Universitäten nicht flächendeckend ab. Es handelt sich nur um eine Stichprobe, die als ausschliessliche Basis eines vollständigen Kompetenzmodells für Studierfähigkeit noch zu eng wäre. Es erwies sich deshalb als notwendig, bereits bestehende Kompetenzmodelle in den Fächern auf ihre Tauglichkeit hinsichtlich der Erhebungsziele von EVAMAR II zu untersuchen, sie zu übernehmen oder sie weiterzuentwickeln. Bei diesen datengestützten Adaptionen ging es im Wesentlichen darum, die eher allgemein und formal gehaltenen Beschreibungen der Bereiche bestehender Kompetenzmodelle anhand des spezifischen Wissens und Könnens, das in Projekt A ermittelt wurde, im Hinblick auf die in EVAMAR II anvisierte Schnittstellenproblematik weiter zu konkretisieren. Die Modellierung erfolgte für die drei untersuchten Fächer Erstsprache, Mathematik und Biologie. Im Sinne der Abgrenzung von dem in der laufenden Diskussion bereits belegten Begriff bzw. Konstrukt des „Kompetenzmodells" mit umfassendem Anspruch werden die Ergebnisse dieser Modellierung als „Kompetenzraster" bezeichnet.

Für alle drei Testfächer entstanden auf diese Weise Kompetenzfelder, welche zum im Projekt A ermittelten Wissen und Können passen. Im Projektteil C wurden dann zu diesen Feldern Aufgaben konstruiert, welche sich auf Fachinhalte aus Projekt A beziehen. Alle Testaufgaben können also sowohl in der gewählten Kompetenzmodellierung als auch in den Lehrmaterialien an Universitäten verortet werden. Diese konkrete und konsequente „Schnittstellen-Verknüpfung" der Tests bzw. der Leistungsmessung im Projekt EVAMAR II wurde noch nie in einer derartigen inhaltlichen Nähe vorgenommen, sie ist also neuartig. Die Kompetenzraster sind im Hauptbericht beschrieben, ihre Komponenten werden weiter unten im Rahmen der Darstellung der Testergebnisse benannt. Die folgenden bereits bestehenden Kompetenzmodelle wurden als am bedeutsamsten für EVAMAR II befunden: der Gemeinsame Europäischen Referenzrahmen für Sprachen (GER), die Modellierung der Kompetenzen für Mathematik in der dritten internationalen Mathematik- und Naturwissenschaftsstudie TIMSS III sowie die Einheitlichen Prüfungsanforderungen in der Abiturprüfung Biologie (EPA) der Kultusministerkonferenz (KMK) in Kombination mit den Kapiteleinteilungen von vier deutschsprachigen Standardlehrwerken der Biologie auf der Gymnasialstufe (Stufe SII). Es ist zu betonen, dass die Kompetenzraster aus der Anlage des Projekts heraus nicht notwendigerweise den gültigen Curricula der drei Testfächer entsprechen. Das gilt insbesondere auch für Erstsprache. Viele curriculare Inhalte dieses Faches, denen aus anderen Gründen als der Vorbereitung auf viele Studienfächer grosse Bedeutung zukommt, sind im Kompetenzraster und dem darauf basierenden Test nicht enthalten.

4 Ergebnisse II: Leistungsmessung – TP C

4.1 Beschreibung der eingesetzten Tests und der Stichprobe

In Rahmen des Teilprojekts C wurden Erhebungen bei Maturandinnen und Maturanden durchgeführt. Sie setzten sich aus Tests in den drei Fachbereichen Erstsprache, Mathematik und Biologie, einer Zusatzbefragung sowie einem überfachlichen Fähigkeitstest (UFT) zusammen. Die Notwendigkeit der Zusatzbefragung und des Einsatzes des UFT ergab sich erst aufgrund der oben beschriebenen Voruntersuchungen (siehe Abschnitte 3.1.2 und 3.1.3). Der Einbezug eines naturwissenschaftlichen Fachs anstatt – wie zuerst vorgeschlagen – Englisch wurde vom Auftraggeber gewünscht. Die konkrete Wahl aus Biologie, Chemie und Physik war zum Zeitpunkt der Auftragserteilung noch offen. Die Steuergruppe entschied sich vor allem deshalb für Biologie, weil dieses Fach im Vergleich zu den anderen beiden für die grössere Zahl von Studienfächern Grundlagen liefert (siehe auch die Ergebnisse der Befragung Dozierender), und weil die Art der Testleistungen sich am ehesten von den bereits in Mathematik und im UFT gemessenen unterscheidet. In der Zusatzbefragung wurde in Ergänzung zum Projekt EVAMAR I auch nach der Selbsteinschätzung weiterer Kompetenzen

im Rahmen der Studierfähigkeit gefragt. In der Deutschschweiz war dieser Fragebogen aufgrund der dort verfügbaren Befragungszeit noch umfassender. Der UFT enthält Items, die in anderen Untersuchungen gut mit dem Studienerfolg in naturwissenschaftlichen Fächern korreliert haben. Er setzt sich aus validierten Items von vier Dimensionen des Eignungstests für das Medizinstudium (EMS) zusammen, die einen besonders hohen Vorhersagewert für Prüfungen in universitären Fachstudien haben. Dieser Zusatztest wurde nur in der Deutschschweiz eingesetzt, weil in der Einschätzung der Projektleitung die Beanspruchung zusätzlicher, in der ursprünglich kommunizierten Planung nicht vorgesehener Testzeit in der Romandie und dem Tessin nicht mehr zumutbar war.

Als Grundlage für die Konstruktion von Aufgaben dienten die im Teilprojekt B erstellten Kompetenzraster, die in der Ergebnisdatenbank der Teilprojekte A1 und A3 gespeicherten Wissens- und Könnenselemente sowie die entsprechenden konkreten Lehrmaterialien. Dabei wurde eine möglichst repräsentative Verteilung auf die Kompetenzfelder einerseits und die ermittelten Wissens- und Könnenselemente andererseits angestrebt. Die Ergebnisse der Befragung von Dozierenden in Teilprojekt A4 erforderten – mit der bereits erwähnten Ausnahme in Erstsprache – keine Korrektur dieser Grundlage. Um die ganze Breite der Kompetenzraster einbeziehen zu können, wurden insgesamt Testaufgaben entwickelt, die von einem Schüler oder einer Schülerin in 720 Minuten gelöst werden könnten. Da für die Erhebungen nicht soviel Zeit zur Verfügung stand, konnte nicht jeder Schülerin bzw. jedem Schüler alle Aufgaben vorgelegt werden. Pro Test konnten lediglich 45 Minuten verfügbar gemacht werden. Für eine zuverlässige Schätzung der Fähigkeiten der Schülerinnen und Schüler, aber auch der Schwierigkeit der Aufgaben, wurde deshalb ein rotierendes Testdesign gewählt, bei dem die Aufgaben untereinander verlinkt sind (Multi-Matrix-Design, bei dem die Testhefte über gleiche Aufgabenblöcke miteinander verknüpft sind, sodass sich das für die einzelne Person ermittelte Wissen und Können nicht nur auf das von ihr bearbeitete Testheft stützen muss). Die Testhefte wurden zufällig auf die Schülerinnen und Schüler verteilt, sodass jede Aufgabe in jeder Klasse jeweils von ähnlich vielen Schülerinnen und Schülern bearbeitet wurde. Bei der Wahl der Aufgabenformen dominierten wie bei vergleichbaren grossen Untersuchungen (TIMSS, PISA) die Kriterien Auswertungsobjektivität und Auswertungsökonomie. Es wurden möglichst wenig Aufgaben zur ausführlichen offenen Beantwortung eingesetzt (am meisten in Erstsprache), weil diese zur Sicherung der Objektivität aufwändige Codierungsarbeiten nach sich gezogen hätten. Die Beschränkung auf die schriftliche Durchführungsform hatte auch eine Einschränkung der testbaren Kompetenzfelder in Erstsprache zur Folge (z. B. Verzicht auf Hör- und Sprechkompetenzen).

Mit der gewählten Kombination von Testinstrumenten konnte insgesamt eine Messung von im Hinblick auf die Studierfähigkeit vielfältigen Aspekten sichergestellt werden (fachliches Wissen, allgemeine Sprachkompetenzen, mathematische Kompetenzen, biologische und allgemeinere naturwissenschaftliche Kompetenzen).

Die Erhebungen wurden in einem über alle Schulen vergleichbaren Zeitraum durchgeführt. Massgebend für den Erhebungszeitpunkt war die Bedingung, dass die Maturandinnen und Maturanden maximal drei Wochen vor Ende des regulären Unterrichts vor den Maturaprüfungen stehen sollten. Wegen der terminlichen Unterschiede zwischen verschiedenen Schulen ergab sich daraus ein Erhebungszeitraum von Ende April bis Anfang Juli 2007.

Zur Grundpopulation gehören die Gymnasiastinnen und Gymnasiasten der Schweiz, die im Sommer 2007 die Maturitätsprüfungen abgelegt haben. Von der Population ausgeschlossen waren die Gymnasiastinnen und Gymnasiasten der Kantone Basel-Landschaft und Genf. Im Kanton Basel-Landschaft findet die Maturitätsprüfung jeweils erst im Herbst statt, weshalb die Vergleichbarkeit nicht gegeben war. Die Gymnasien des Kantons Genf erhoben schon

früh grosse Einwände gegen die Untersuchung, und die meisten Schulen verweigerten die Übermittlung der Angaben, die Ende 2006 eingeholt und zur Bestimmung der Grundgesamtheit notwendig gewesen wären. Die Steuergruppe des Projekts EVAMAR II verzichtete in der Folge aus verschiedenen Gründen darauf, eine allfällige, von den zuständigen Behörden von oben verordnete Teilnahme zu erwirken. Eine zwangsweise Teilnahme hätte ein erhebliches Verfälschungspotenzial für die Validität der Ergebnisse mit sich gebracht. Um verschiedene Vergleiche von Subpopulationen durchführen zu können, wurde die Stichprobe stratifiziert. Nicht alle gewünschten Klassen haben an den Erhebungen teilgenommen. Problematisch ist aber einzig die aus verschiedenen Gründen niedrig ausgefallene Beteiligungsquote der Tessiner Schulen (siehe im Einzelnen im Hauptbericht). Deshalb konnten die Ergebnisse des Tessins in diesem Teilprojekt nicht in die Auswertungen einbezogen werden. Die Rücklaufquote ohne Tessin beträgt insgesamt 91% bei den Klassen und 85% bei den Personen. Die tiefere zweite Zahl erklärt sich aus Absenzen an den Erhebungstagen. Insgesamt lagen auswertbare Daten von rund 3'800 Personen vor.

4.2 Gesamtergebnisse

Es folgt eine Darstellung der Ergebnisse der Gesamtstichprobe. Für die Testergebnisse werden dabei zwei Berechnungsvarianten vorgelegt:

- Die Berechnungsvariante A dient der Veranschaulichung: Die Werte für die einzelnen Themen- und Kompetenzbereiche ergeben sich aus der Addition der Ergebnisse in den einzelnen, in die Hauptauswertung einbezogenen Aufgaben. Damit soll insbesondere auch jeweils der Anteil richtiger Lösungen gemessen am Total der möglichen richtigen Lösungen aufgezeigt werden. Weil jeweils ein einzelner Schüler nur zwei von mehreren Blöcken mit jeweils unterschiedlichem Umfang an Aufgaben bearbeitet hat, wurden die Ergebnisse der Fachtests für jedes Heft pro Bereich auf jeweils maximal 20 Punkte (bzw. 10 Punkte bei Biologie) standardisiert. Dass die Teilbereiche in den verschiedenen Blöcken unterschiedlich schwer sind, spielt bei der Darstellung der nationalen Gesamtergebnisse insofern keine Rolle, als die Hefte nach Zufall gleichverteilt waren.
- Berechnungsvariante P ist die noch genauere Schätzung der Personenfähigkeiten nach dem Rasch-Modell. Ein Wert von 0 entspricht einer mittleren Personenfähigkeit. Die Skala ist grundsätzlich nach unten und nach oben offen, die Werte liegen aber i.d.R. zwischen -3 und +3. Die geschätzte Personenfähigkeit stützt sich auf alle aus den verschiedenen Testheften einbezogenen Aufgaben und berücksichtigt zudem ihren unterschiedlichen Schwierigkeitsgrad.

Tab. 2: Gesamtergebnisse Erstsprache

Bereiche	Mittel A	S A	Max A	Mittel P	S P	Min P	Max P
TOT_SPRACHE	31.22	9.27	60.00	.20	.562	-1.84	2.03
ALLG_LV	9.89	4.94	20.00	.38	.621	-2.11	2.30
DETAIL_LV	10.52	3.97	20.00	.44	.676	-2.29	2.31
WORTS	10.81	4.39	20.00	.39	.424	-1.18	1.90
GRAM	13.49	2.60	18.92	.38	.399	-1.36	2.32

In Tabelle 2 sind die Gesamtergebnisse in Erstsprache aufgeführt. Die Abkürzungen in der Tabelle verstehen sich wie folgt:

- TOT_SPRACHE: Gesamtergebnis Erstsprache (ohne Grammatik)
- ALLG_LV: Allgemeines Leseverstehen (zur Orientierung lesen)
- DETAIL_LV: Detailliertes Leseverstehen (Information & Argumentation verstehen)
- WORTS: Sprachreflexion/Wortschatz
- GRAM: Sprachreflexion/Grammatische Kompetenz und Orthografie
- Mittel A: Mittelwerte nach Variante A (max. 20 Punkte pro Bereich, max. 60 Punkte total)
- S A: Standardabweichung zu Mittel A

- Max A: Erreichte Maximalpunktzahl Variante A
- Mittel P: Mittelwerte nach Variante P (Wert null entspricht einer „mittleren“ Fähigkeit)
- S P: Standardabweichung zu Mittel P
- Min P: Tiefste Personenfähigkeit Variante P
- Max P: Höchste Personenfähigkeit Variante P

Der aus konstruktionstechnischen Gründen nur in der Deutschschweiz eingesetzte Teil „Grammatische Kompetenz und Orthografie“ des Erstsprachtests ist im Gesamttotal nicht enthalten und wird separat ausgewiesen. Bei der Berechnungsvariante P ermöglicht die Verlinkung der Testhefte Fähigkeitsschätzungen für diesen Bereich auch für die Romandie.

Im Mittel wurden die Aufgaben zu etwas mehr als der Hälfte richtig gelöst (im Bereich „Grammatische Kompetenz und Orthografie“ deutlich mehr). Die geschätzten Personenfähigkeiten liegen im Durchschnitt in allen Kompetenzbereichen auf einem Niveau, auf dem die Maturandinnen und Maturanden Aufgaben lösen konnten, deren Anforderungen über einem mittleren Schwierigkeitsgrad liegen. Die Streuung auf Ebene der einzelnen Personen ist beachtlich. Mit Ausnahme des Bereichs „Grammatische Kompetenz und Orthografie“ wurde jeweils auch die maximal mögliche Punktzahl erreicht. Eine zusätzliche Auswertung der über die Klassen aggregierten Werte ergibt eine erhebliche Streuung auch zwischen den Klassenmittelwerten (Minima: TOT_SPRACHE = 20.72, GRAM = 7.55; Maxima: TOT_SPRACHE = 41.41, GRAM = 17.23). Die beste Klasse hat im Mittel doppelt so viele Aufgaben richtig gelöst wie die schlechteste.

Tab. 3: Gesamtergebnisse Mathematik

Bereiche	Mittel A	S A	Max A	Mittel P	S P	Min P	Max P
TOT_MATH	37.65	13.25	80.00	-.09	.721	-2.22	2.44
ANALYSIS	7.79	4.71	20.00	-.12	.896	-2.70	2.90
GEOMETRIE	9.08	4.58	20.00	-.08	.774	-2.28	2.49
STOCHASTIK	11.02	5.12	20.00	-.08	.990	-3.46	3.24
M_VERSCH	9.76	4.76	20.00	-.11	.818	-2.55	2.64

In Tabelle 3 sind die Gesamtergebnisse in Mathematik aufgeführt. Die neu verwendeten Abkürzungen in der Tabelle bedeuten Folgendes:
- TOT_MATH: Gesamtergebnis Mathematik
- ANALYSIS: Analytisches (Funktionen, Infinitesimalrechnung, Folgen und Reihen)
- GEOMETRIE: Geometrisches (Trigonometrie, analytische Geometrie, Vektorrechnung)
- STOCHASTIK: Stochastisches (Wahrscheinlichkeitsrechnung, Kombinatorik, Statistik)
- M_VERSCH: Diverses (Mengenlehre/diskrete Mathematik, Logik, elementare Algebra)
- Mittel A: Mittelwerte nach Variante A (max. 20 Punkte pro Bereich, max. 80 Punkte total)

Im Mittel wurden die Aufgaben von den Testpersonen zu etwas weniger als der Hälfte richtig gelöst, die Analysisaufgaben zu rund 40%, die Geometrieaufgaben zu 45%, die Stochastikaufgaben zu 55% und die Aufgaben des übrigen Bereichs zu 50%. Werden die Aufgabenschwierigkeiten in die Schätzung der Personenfähigkeiten miteinbezogen, nähern sich die Durchschnittswerte einander an, insbesondere auch die Bereiche Analysis (gegen oben) und Stochastik (gegen unten). Das eher schlechte Rohergebnis in Analysis muss deshalb insofern relativiert werden, als diese Aufgaben besonders schwierig waren. Insgesamt liegen die geschätzten Personenfähigkeiten im Durchschnitt in allen Kompetenzbereichen auf einem Niveau, auf dem die Maturandinnen und Maturanden Aufgaben lösen konnten, deren Anforderungen ganz leicht unter einem mittleren Schwierigkeitsgrad liegen. Die Streuung zwischen den einzelnen Personen ist wieder erheblich. In allen Bereichen wurde auch das Punktemaximum erreicht. Eine zusätzliche Auswertung der über die Klassen aggregierten Werte ergibt wieder eine bemerkenswerte Streuung auch zwischen den Klassenmittelwerten

(Minimum: TOT_MATH = 22.36; Maximum: TOT_MATH = 64.04). Die beste Klasse hat also im Mittel beinahe dreimal so viele Aufgaben richtig gelöst wie die schlechteste.

In Tabelle 4 sind die Gesamtergebnisse in Biologie aufgeführt. Da an vielen Gymnasien dieses Fach bereits ein bis zwei Jahre vor den Maturaprüfungen abgeschlossen wird, werden die Resultate für zwei Gruppen („abgeschlossen ja", „abgeschlossen nein") dargestellt.

Tab. 4: Gesamtergebnisse Biologie

	Mittel A		S A		Max A		Mittel P		S P		Min P		Max P	
abgeschl.	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein
TOT_BIO	26.8	32.5	8.71	8.84	54.2	56.3	-.22	.10	.470	.474	-1.69	-1.28	1.24	1.59
B_EVO	4.79	5.45	3.00	2.94	10.0	10.0	-.21	.09	.452	.472	-1.68	-1.68	1.21	1.35
B_STRU	3.85	4.75	2.55	2.79	10.0	10.0	-.22	.10	.460	.484	-2.07	-1.80	1.23	1.56
B_GEN	4.44	5.42	2.38	2.39	10.0	10.0	-.21	.08	.410	.428	-1.81	-1.70	1.19	1.52
B_INFO	4.16	5.28	2.46	2.48	10.0	10.0	-.24	.11	.497	.514	-1.97	-1.91	1.30	1.45
B_OEKO	4.85	5.77	2.63	2.60	10.0	10.0	-.24	.12	.533	.539	-2.22	-1.87	1.33	1.61
B_STOFF	4.56	5.78	2.53	2.56	10.0	10.0	-.25	.14	.552	.569	-2.40	-2.07	1.52	1.76

Die neu verwendeten Abkürzungen in der Tabelle verstehen sich wie folgt:
- TOT_BIO: Gesamtpunktzahl
- B_EVO: Evolution (Entwicklungsprozesse – Evolution und Zukunftsfragen)
- B_STRU: Struktur und Funktion (Bau und Funktion von Zellen, Geweben und Organen; funktionsbezogene Differenzierungen)
- B_GEN: Genetik (Grundlagen der molekularen Genetik, Anwendung moderner biologischer Erkenntnisse und Methoden)
- B_INFO: Informationsverarbeitung (Kommunikation zwischen Zellen, neuronale Informationsverarbeitung, Wahrnehmung)
- B_OEKO: Ökologie (Vernetzte Systeme – Ökologie und Nachhaltigkeit)
- B_STOFF: Stoffwechsel (Stoff- und Energiewechsel: Notwendigkeit und Wege der Energieumwandlung, Assimilation und Dissimilation im Zusammenhang von zellulären Strukturen und Organismus; Molekulare Steuerung von Stoffwechsel: Enzymatik)
- Mittel A: Mittelwerte nach Variante A (max. 10 Punkte pro Bereich, max. 60 Punkte total)

Im Mittel wurden die Testaufgaben in Biologie knapp zur Hälfte richtig gelöst, wobei die Gruppe mit zurückliegendem Fachabschluss noch schlechter, die andere Gruppe hingegen überdurchschnittlich abgeschlossen hat. Die Biologieaufgaben wurden von der ersten Gruppe auch im Vergleich zu Erstsprache und Mathematik eher schlechter gelöst. D. h., auch die geschätzten Personenfähigkeiten liegen im Durchschnitt unter einem mittleren Fähigkeitsmass. Vermutlich spielt hier der Vergessenseffekt eine Rolle, der aber auch weit höher hätte ausfallen können. Im Gesamttotal wurde die maximal mögliche Punktzahl von niemandem erreicht, wenn zum Teil auch nur knapp verfehlt. Die Streuung auf Ebene der einzelnen Personen ist wieder beachtlich. In jedem Bereich wurde jeweils auch die maximal mögliche Punktzahl erreicht. Die zusätzliche Auswertung der über die Klassen aggregierten Werte ergibt nochmals eine erhebliche Streuung auch zwischen den Klassenmittelwerten (Minimum: B_TOTAL_B2H = 14.93; Maximum: B_TOTAL_B2H = 42.35). Wieder hat die beste Klasse im Mittel beinahe dreimal so viele Aufgaben richtig gelöst wie die schlechteste.

Die Auswertung der Testergebnisse offenbart in allen Bereichen eine erstaunlich breite Streuung, vor allem angesichts der Tatsache, dass die Schülerinnen und Schüler kurz vor der Verleihung der für alle Studienfächer geltenden, universalen Qualifikation „Hochschulreife" standen. Es kann deshalb davon ausgegangen werden, dass nicht alle Maturandinnen und Maturanden in allen drei getesteten Fachbereichen über Kompetenzen verfügen, die den

universitären Anforderungen aller Studienfächer entsprechen[1]. Deshalb soll zum Vergleich auch ein Blick auf die Mittelwerte und die Streuungen effektiv vergebener Maturanoten in den drei Testfächern geworfen werden (siehe Tabelle 5).

Tab. 5: Ausgewählte Maturanoten in den drei Testfächern

Prüfungen	Min	Max	Mittel	S
Erstsprache schriftlich	2.00	6.00	4.41	.739
Erstsprache Gesamtnote	2.50	6.00	4.63	.542
Mathematik schriftlich	1.00	6.00	4.03	1.048
Mathematik Gesamtnote	1.50	6.00	4.34	.826
Biologie Erfahrungsnote	2.50	6.00	4.68	.541
Naturwissenschaften Gesamtnote	2.50	6.00	4.60	.566

Ausser in den schriftlichen Mathematikprüfungen wurde der minimal mögliche Wert (Note 1) nicht vergeben. Die relative Streuung fällt insgesamt kleiner aus als bei den Tests. Es fällt zudem auf, dass die Streuung in Mathematik erheblich grösser ist als in den beiden anderen Fächern. Ebenso, dass in den schriftlichen Mathematikprüfungen die Durchschnittsnote bei vier, also gerade noch bei „genügend“ liegt. Eine Auswertung der über die Klassen aggregierten Zahlen ergibt folgende Extremwerte der Klassendurchschnitte: Minima: Erstsprache schriftlich 3.6, Erstsprache Gesamtnote 3.97, Mathematik schriftlich 2.88, Mathematik Gesamtnote 3.73, Biologie Erfahrungsnote 3.85, Naturwissenschaften Gesamtnote 4.0; Maxima: Erstsprache schriftlich 5.11, Erstsprache Gesamtnote 5.17, Mathematik schriftlich 5.48, Mathematik Gesamtnote 5.44, Biologie Erfahrungsnote 5.28, Naturwissenschaften Gesamtnote 5.31.

Bei Notenbewertungen steht fest, dass Noten unter 4.0 als ungenügende Qualifikation gelten. Demgemäss können aus einer weiteren Auswertung der Notendaten die nachfolgend aufgeführten Prozentanteile von Maturandinnen und Maturanden als in einzelnen Bereichen von den Schulen selbst als ungenügend qualifiziert angegeben werden (Note 3.9 oder tiefer). Schriftliche Erstsprachprüfung: 19.6%, Gesamt-Maturanote in Erstsprache: 4.7%, Schriftliche Mathematikprüfung: 41.4%, Gesamt-Maturanote in Mathematik: 24.4%, Biologie-Erfahrungsnote: 5.6%, Gesamt-Maturanote in Naturwissenschaften: 5.6%. In der schriftlichen Maturaprüfung in Mathematik wurden also 41.4% der Schweizer Maturandinnen und Maturanden als ungenügend bewertet. 19.6% erzielten in der schriftlichen Erstsprachprüfung (überwiegend in Form eines Aufsatzes) ungenügende Noten. Die an der Maturaprüfung mittels schriftlicher Leistungsmessung festgestellten Kompetenzen in diesen Fächern wurden somit durch die Schulen selbst bzw. durch die entsprechenden Lehrpersonen beim durch diese Zahlen wiedergegebenen Anteil von Maturandinnen und Maturanden als ungenügend bewertet. Die Zahlen steigen zudem beim Übergang in den genügenden Bewertungsbereich (Note 4 oder tiefer) sprunghaft an. Vermutlich neigt man bei dieser Grenze zum Aufrunden, sodass die objektive Zahl der ungenügenden Qualifikationen eher noch höher ausfallen dürfte. Die oben aufgrund der Leistungstestergebnisse von EVAMAR II angebrachte Vermutung partiell ungenügender Kompetenzen bestätigt deshalb eigentlich – zumindest teilweise – nur eine auch von den Examinatorinnen und Examinatoren ausgeübte Bewertungspraxis. Einzig die genauen Zahlen fallen anders aus. Bei der Notengebung dürfte zudem auch eine gewisse Anpassung an die Leistungsfähigkeit von Klassen eine Rolle spielen.

In Tabelle 6 sind die Gesamtergebnisse für die vier Bereiche des UFT aufgeführt. Die neu verwendeten Abkürzungen in der Tabelle bedeuten Folgendes:
- U_QUANT: Quantitative und formale Probleme

[1] Eine genauere Analyse auf der Ebene von konkreten Aufgaben und Teilbereichen erfolgt im Hauptbericht.

- U_DIAGR: Diagramme und Tabellen
- U_TEXT: Textverständnis
- U_NAT: Naturwissenschaftliches Grundverständnis
- Mittel A: Erreichte Maximalpunktzahl Variante A (max. 10 Punkte pro Bereich, 12 Punkte bei U_TEXT)

Tab. 6: Gesamtergebnisse Überfachlicher Fähigkeitstest (UFT)

Bereiche	Mittel A	S A	Max A	Mittel P	S P	Min P	Max P
U_QUANT	4.57	2.496	10.00	.06	.839	-2.93	2.55
U_DIAGR	5.28	2.270	10.00	.08	.779	-2.37	2.33
U_TEXT	5.75	2.512	12.00	.09	.717	-2.39	2.32
U_NAT	4.58	2.191	10.00	.04	.712	-2.39	2.41

Im Mittel wurden die Aufgaben des UFT in den Bereichen „Quantitative und formale Probleme“, „Textverständnis“ und „Naturwissenschaftliches Grundverständnis“ zu etwas weniger als der Hälfte richtig gelöst. Einzig im Bereich „Diagramme und Tabellen“ liegen die Werte etwas über der Hälfte. Insgesamt befinden sich die Werte leicht unter den Ergebnissen des vergleichbaren Eignungstests für das Medizinstudium (EMS). Der tiefere Wert lässt sich damit erklären, dass bei Weitem nicht alle Maturandinnen und Maturanden den EMS ablegen und die entsprechende freiwillige Vorselektion bereits zu besseren Ergebnissen führt. Diese Vergleichswerte deuten auch darauf hin, dass sich die Probanden offenbar angestrengt haben. Die Ergebnisse im UFT sind somit auch als ein wichtiger Indikator für eine überwiegend seriöse Bearbeitung aller Tests zu deuten. Die geschätzten Personenfähigkeiten liegen im Durchschnitt in allen Kompetenzbereichen auf einem Niveau, auf dem die Maturandinnen und Maturanden Aufgaben lösen können, deren Anforderungen leicht über einem mittleren Schwierigkeitsgrad liegen. Auffällig ist wieder die breite Streuung. In allen Bereichen wurde von einzelnen Maturandinnen und Maturanden auch das Punktemaximum erreicht. Eine zusätzliche Auswertung der über die Klassen aggregierten Werte ergibt die im Vergleich zu den anderen Tests grössten Differenzen zwischen den Mittelwerten der jeweils besten und schlechtesten Klasse (Minima: U_QUANT=1.60, U_DIAGR=2.50, U_TEXT=2.44, U_NAT=2.25; Maxima: U_QUANT=8.30, U_DIAGR=8.40, U_TEXT=9.56, U_NAT=7.67).

In die ergänzende Zusatzbefragung wurde die Selbsteinschätzung folgender weiterer überfachlicher Kompetenzen aufgenommen, die (auch gemäss Befragung der Dozierenden) für den Studienalltag wichtig sind: Zeitplanung, zeiteffiziente Prüfungsvorbereitung, Zeiteffizienz, Konzentrationsfähigkeit, vernetztes Denken, Diskussionsfähigkeit, Perspektivenübernahme. Die Maturandinnen und Maturanden sollten auf einer Skala von 1 (= sehr selten) bis 5 bzw. 4 (=sehr oft) angeben, inwiefern Aussagen, welche die genannten überfachlichen Kompetenzen thematisieren, auf sie selbst zutreffen. In Tabelle 7 finden sich die Mittelwerte und Standardabweichungen der verschiedenen Skalen.

Tab. 7: Mittelwerte der Skalen der erfragten überfachlichen Kompetenzen

Kompetenzbereiche	Min	Max	Mittel	S
Zeitplanung	1.00	5.00	2.64	.961
Effiziente Prüfungsvorbereitung	1.00	5.00	3.66	.666
Zeiteffizienz	1.00	5.00	3.76	.646
Konzentrationsfähigkeit	1.00	5.00	2.86	.700
Vernetztes Denken	1.00	5.00	3.29	.777
Diskussionsfähigkeit	1.00	4.00	3.28	.513
Perspektivenübernahme	1.00	4.00	2.99	.578

Es lässt sich ablesen, dass die selbst berichteten überfachlichen Kompetenzen unterschiedlich stark ausgeprägt sind. Während die Maturandinnen und Maturanden ihrer Einschätzung zufolge „selten bis manchmal“ aktive Zeitplanung betreiben (M=2.64) und auch angeben, sich nur „selten bis manchmal“ auf eine Sache konzentrieren zu können (M=2.86), haben sie offensichtlich wenig Schwierigkeiten, sich effizient auf Prüfungen vorzubereiten (M=3.66) oder ihre Zeit effizient zu nutzen (M=3.76).

4.3 Vergleiche

Im heterogenen System des Schweizer Gymnasiums sind Vergleiche zwischen verschiedenen Gruppen von Gymnasien sowie Maturandinnen und Maturanden von besonderem Interesse. Diese werden im Folgenden angestellt. Dazu werden ausschliesslich die aus der „Rasch-Modellierung“ berechneten Personenfähigkeiten verwendet, die sich jeweils auf alle Testhefte beziehen (Darstellungsvariante P). Im Unterschied zur bisherigen Berechnung der Personenfähigkeiten sind diese nun jeweils zusätzlich auf eine Skala von 200-800 Punkten und einen gesamtschweizerischen Mittelwert von 500 standardisiert. Im Bereich zwischen 400 und 600 Punkten liegen etwa zwei Drittel (68.2 Prozent) aller Maturandinnen und Maturanden, zwischen 300 und 700 Punkten 95.4 Prozent, und nur 2.3 Prozent erreichen weniger als 300 oder mehr als 700 Punkte. Durch diese Standardisierung können Unterschiede über verschiedene Tests hinweg besser vergleichbar gemacht werden.

Für die Vergleiche von mehr als zwei Gruppen wurde jeweils ein a-posteriori-Test für multiple Mittelwertvergleiche auf Signifikanz nach Duncan durchgeführt. Bei nur zwei Gruppen wurde die Signifikanz der Unterschiede mit einem t-Test untersucht. Wegen der grossen Stichprobenumfänge können auch kleine Mittelwertsunterschiede mit wenig Praxisrelevanz signifikant sein. Deshalb wurde für die Vergleiche das Mass „d“ der Effektstärke nach Cohen berechnet, und zwar für die jeweils grösste Mittelwertsdifferenz. Das Effektstärkenmass wird häufig wie folgt interpretiert: d=0.20: kleiner Effekt; d=0.50: mittlerer Effekt; d=0.80: grosser Effekt. Die so errechneten Werte sind in den folgenden Tabellen zusammengestellt. Die Zahl in Klammern bezeichnet jeweils die Zugehörigkeit zu Gruppen, die sich signifikant unterscheiden. Die Zahl „(1)“ bezeichnet dabei die Gruppe mit den niedrigsten Werten. Die Angabe von zwei Zahlen steht für das Fehlen signifikanter Unterschiede zu den anderen Mitgliedern beider Gruppen. **Tiefere Testwerte dürfen keinesfalls mit schlechterer Unterrichtsqualität an den entsprechenden Gymnasien gleichgesetzt werden, sondern haben i.d.R. strukturelle Ursachen!** Die in den Tabellen verwendeten Abkürzungen für die Testbereiche wurden bereits in den bisherigen Tabellen erläutert.

4.3.1 Vergleiche zwischen den Straten (Gruppen von Gymnasien)

Die Vergleiche zwischen den verschiedenen Straten finden sich in den Tabelle 8 bis 11. Die Strateneinteilung erfolgte entsprechend folgender Überlegungen: Zürich ist grösster Kanton; Der deutschsprachige Teil des Kantons Bern ist jene Deutschschweizer Region, in der bis zum Maturitätsabschluss – als strukturelle Eigenheit – das erste Jahr von vier Jahren an einer Sekundarschule stattfinden kann und nur drei Jahre im eigentlichen Gymnasium absolviert werden müssen („Deutschschweiz MD3“); „Deutschschweiz klein“ umfasst die Kantone mit weniger als 15 Maturaklassen (AI, NW, OW, GL, UR, AR, SH, VSd, ZG); „Deutschschweiz gross“ enthält die Kantone mit mehr als 15 Klassen (SZ, SO, TG, GR, BS, AG, SG, LU); „Romandie 1“ besteht aus jenen Westschweizer Kantonen, in denen ebenfalls nur drei Jahre zwingend am eigentlichen Gymnasium absolviert werden müssen (BEf, JU, NE, VD); und „Romandie 2“ umfasst die Westschweizer Kantone mit mindestens vier Jahren Ausbildungsdauer am eigentlichen Gymnasium (FR, VSf).

Tab. 8: Vergleich der Testergebnisse in Erstsprache zwischen den Straten

Testbereich	Zürich	Deutsch-schweiz MD3	Deutsch-schweiz klein	Deutsch-schweiz gross	Romandie 1	Romandie 2	grösste Effekt-stärke
TOT_SPRACHE	498 (2)	489 (1)	502 (2)	501 (2)	501 (2)	524 (3)	0.41
ALLG_LV	500 (1)	495 (1)	507 (2)	498 (1)	499 (1)	514 (3)	0.24
DETAIL_LV	502 (3)	487 (1)	510 (4)	504 (3,4)	493 (2)	519 (5)	0.38
WORTS	499 (2)	492 (1)	501 (2)	493 (1)	512 (3)	517 (3)	0.34
GRAM	502 (2)	489 (1)	509 (3)	500 (2)	500 (2)	512 (3)	0.32

In Erstsprache (siehe Tabelle 8) schnitten die Deutschschweizer Gymnasien mit drei Jahren Mindestdauer im Mittel signifikant am schlechtesten ab. Die Effektstärken zur jeweils besten Gruppe, Romandie 2, liegen zwischen gering und mittel. Auf der Ebene der einzelnen Kompetenzbereiche ergeben sich die folgenden Differenzierungen: Beim „Allgemeinen Leseverstehen" schnitten nur noch die kleinen Deutschschweizer Kantone und die Gruppe „Romandie 2" besser ab als die Straten mit den tiefsten Werten. Beim „Detaillierten Leseverstehen" ergibt sich eine nochmals differenziertere Gruppierung. Hier bildet auch das Stratum „Romandie 1" eine eigene Gruppe am unteren Ende. Im Bereich „Sprachreflexion/Wortschatz" fallen auch die grossen Deutschschweizer Kantone in die unterste Gruppe. Die Verteilung im Bereich „Grammatik" bestätigt im Wesentlichen das Bild des Gesamtergebnisses. Allerdings gehören hier auch die kleinen Deutschschweizer Kantone zur Gruppe der Besten. Insgesamt schnitten die Gymnasien mit nur dreijähriger Mindest-Dauer zumindest innerhalb des gleichen Sprachbereichs eher schlechter ab.

Tab. 9: Vergleich der Testergebnisse in Mathematik zwischen den Straten

Testbereiche	Zürich	Deutsch-schweiz MD3	Deutsch-schweiz klein	Deutsch-schweiz gross	Romandie 1	Romandie 2	grösste Effekt-stärke
TOT_MATH	501 (3,4)	492 (2)	505 (4)	496 (2,3)	477 (1)	522 (5)	0.48
ANALYSIS	498 (3)	492 (2)	505 (4)	496 (2,3)	483 (1)	520 (5)	0.41
GEOMETRIE	501 (3)	498 (2,3)	503 (3)	494 (2)	488 (1)	514 (4)	0.28
STOCHASTIK	498 (3)	492 (2)	500 (3)	501 (3)	481 (1)	528 (4)	0.55
VERSCH	505 (4)	489 (2)	498 (3)	500 (3,4)	476 (1)	517 (5)	0.47

In Mathematik (siehe Tabelle 9) schnitten die Gymnasien mit nur dreijähriger Mindest-Dauer im Gesamtergebnis klar schlechter ab als jene mit mindestens vier Jahren, insbesondere in der Romandie. Die Effektstärken zwischen den Straten mit dem tiefsten (Romandie 1) und dem höchsten Wert (Romandie 2) sind mittlere. Dieses Bild wird insbesondere in den Bereichen „Stochastik" und „Verschiedenes" mit noch leicht höheren Effektstärken akzentuiert und umgekehrt im Bereich „Geometrie" etwas entschärft. Es scheint sich zu zeigen, dass die Schuldauer im Gymnasium in Mathematik eine noch grössere positive Auswirkung auf den Leistungsstand hat als in Erstsprache.

In Biologie (siehe Tabelle 10) liegt Zürich klar an der Spitze, mit einer hohen Effektstärke im Vergleich zum Stratum mit dem tiefsten Wert. Am unteren Ende gruppieren sich die Schulen aus der Romandie. Die Unterschiede zwischen den Gymnasien mit drei- und vierjähriger Mindest-Dauer am Schulort Gymnasium sind innerhalb der Romandie unbedeutend, in der Deutschschweiz hingegen nach wie vor vorhanden, und zwar zugunsten einer vierjährigen Mindest-Dauer.

Tab. 10: Vergleich der Testergebnisse in Biologie zwischen den Straten

Testbereiche	Zürich	Deutsch-schweiz MD3	Deutsch-schweiz klein	Deutsch-schweiz gross	Romandie 1	Romandie 2	grösste Effekt-stärke
TOT_BIO	522 (4)	506 (2)	510 (2,3)	514 (3)	459 (1)	462 (1)	0.75
B_EVO	519 (4)	505 (2)	512 (3)	514 (3,4)	462 (1)	464 (1)	0.74
B_STRU	521 (4)	505 (2)	509 (2)	515 (3)	463 (1)	463 (1)	0.74
B_GEN	521 (4)	505 (2)	509 (2,3)	513 (3)	464 (1)	463 (1)	0.76
B_INFO	522 (4)	504 (2)	509 (3)	513 (3)	462 (1)	463 (1)	0.75
B_OEKO	521 (4)	506 (2)	511 (2,3)	513 (3)	460 (1)	463 (1)	0.75
B_STOFF	522 (4)	505 (2)	510 (2,3)	514 (3)	461 (1)	462 (1)	0.75

Auch im überfachlichen Fähigkeitstest (siehe Tabelle 11) schneidet die Gruppe der Gymnasien mit dreijähriger Mindest-Dauer in allen Unterbereichen signifikant schlechter ab. Die Effektstärken zwischen den Extremgruppen liegen allerdings auf einem tiefen Niveau. Vermutlich bestätigt sich hier der Anspruch des Tests, Fähigkeiten zu prüfen, die relativ unabhängig vom Umfang des in den verschiedenen Fächern erarbeiteten Wissens sind. Die Mindestdauer am Schulort Gymnasium scheint aber trotzdem bedeutsam zu bleiben, wenn auch in geringerem Ausmass.

Tab. 11: Vergleich der Testergebnisse im UFT zwischen den Straten

Testbereich	Zürich	Deutsch-schweiz MD3	Deutsch-schweiz klein	Deutsch-schweiz gross	grösste Effekt-stärke
U_QUANT	506 (3)	487 (1)	502 (3)	496 (2)	0.23
U_DIAGR	507 (4)	487 (1)	501 (3)	495 (2)	0.25
U_TEXT	507 (3)	486 (1)	501 (2)	496 (2)	0.24
U_NAT	507 (3)	486 (1)	500 (2)	496 (2)	0.25

Die in den Vergleichen der Fachtestergebnisse zwischen den Straten festgestellten mittleren bis grossen Unterschiede werfen die Frage auf, ob sich solche auch in den effektiv vergebenen Maturanoten in den drei Fächern zeigen (siehe Tabelle 12).

Tab. 12: Vergleich der Maturanoten in den Testfächern zwischen den Straten

Fächer	Zürich	Deutsch-schweiz MD3	Deutsch-schweiz klein	Deutsch-schweiz gross	Romandie 1	Romandie 2	grösste Effekt-stärke
Erstsprache	4.73 (5)	4.63 (3)	4.76 (5)	4.67 (4)	4.48 (1)	4.55 (2)	0.50
Mathematik	4.39 (2)	4.30 (1)	4.39 (2)	4.35 (1)	4.29 (1)	4.38 (2)	0.12
Biologie	4.58 (1)	4.64 (2)	4.63 (2)	4.73 (3)	4.77 (4)	4.70 (3)	0.36
Naturwiss.	4.54 (1)	4.60 (2)	4.61 (2)	4.66 (3)	4.52 (1)	4.63 (2,3)	0.26

Die Maturanoten in Erstsprache unterscheiden sich zwischen den Straten ebenfalls signifikant, die Effektstärken zwischen den Extremwerten sind mittelgross. Die Rangierung ist aber im Vergleich zu jener bei den Testergebnissen in Erstsprache eine andere. Eine plausible Erklärung findet sich nicht so leicht. Sicher gehört zu den möglichen Gründen, dass ein gymnasiales Erstsprache-Curriculum weit mehr umfasst und z. T. anderes beinhaltet als mit den Tests gemessen wurde. Es ist augenfällig, dass in der Westschweiz über die Sprachgrenzen hinweg in Erstsprache tiefere Noten vergeben wurden als in der übrigen Schweiz. Die Differenz bei der Maturanote in Mathematik zwischen dem tiefsten und höchsten Mittelwert ist zwar signifikant, beträgt aber nur eine Zehntelnote. Die Effektstärke ist hier sehr gering. Interessant ist die im Vergleich zu den Testergebnissen umgekehrte Reihenfolge der besten und schlechtesten Straten bei der Erfahrungsnote Biologie mit immerhin einer beinahe mittleren Effektstärke. Auch bei der Maturanote in Naturwissenschaften ergeben sich zwischen den Straten nur geringfügige Unterschiede. Insgesamt kann festgestellt werden, dass

sich bei den Maturanoten keine oder nur noch kleine Unterschiede zwischen den Straten finden lassen. Vermutlich lässt sich dieser Befund damit erklären, dass Bewertungen an der Matura nicht an einer gesamtschweizerischen Norm erfolgen (können).

4.3.2 Vergleiche nach Schwerpunktfächern

Im Folgenden werden die Testergebnisse der Maturandinnen und Maturanden mit verschiedenen Schwerpunktfächern miteinander verglichen (siehe Tabellen 13 bis 16).

Tab. 13: Vergleich der Testergebnisse in Erstsprache nach gewählten Schwerpunktfächern

Testbereiche	Alte Sprachen	Moderne Sprachen	Physik und Anwendungen der Mathematik	Biologie und Chemie	Wirtschaft und Recht	Philosophie/ Pädagogik/ Psychologie	Bildnerisches Gestalten	Musik	grösste Effektstärke
TOT_SPRACHE	533 (5)	502 (3)	501 (3)	510 (4)	492 (2)	514 (4)	477 (1)	494 (2)	0.59
ALLG_LV	536 (4)	501 (2)	501 (2)	510 (3)	489 (1)	508 (3)	484 (1)	495 (2)	0.63
DETAIL_LV	526 (6)	506 (4,5)	498 (2,3)	503 (3,4)	492 (2)	511 (5)	485 (1)	493 (2)	0.46
WORTS	527 (5)	502 (3)	506 (3,4)	511 (4)	491 (2)	510 (4)	485 (1)	484 (1)	0.58
GRAM	531 (5)	509 (4)	501 (3)	494 (2)	491 (1,2)	511 (4)	487 (1)	488 (1)	0.59

Maturandinnen und Maturanden mit dem Schwerpunktfach „Alte Sprachen“ (vorwiegend Latein) haben im Erstsprachtest in allen Bereichen durchschnittlich am besten, jene mit dem Schwerpunktfach „Bildnerisches Gestalten“ ebenfalls in allen Bereichen und teilweise mit den Gruppen „Musik“ sowie „Wirtschaft und Recht“ am schlechtesten abgeschlossen (siehe Tabelle 13). Die Effektstärken zwischen den Extremgruppen sind mittel. Es fällt auf, dass sich die Schülerinnen und Schüler des Sprachen-Schwerpunkts „Moderne Sprachen“ in einem Sprachtest „nur“ in der Mitte platzierten.

Tab. 14: Vergleich der Testergebnisse in Mathematik nach gewählten Schwerpunktfächern

Testbereiche	Alte Sprachen	Moderne Sprachen	Physik und Anwendungen der Mathematik	Biologie und Chemie	Wirtschaft und Recht	Philosophie/ Pädagogik/ Psychologie	Bildnerisches Gestalten	Musik	grösste Effektstärke
TOT_MATH	512 (6)	475 (3)	614 (7)	504 (5)	485 (4)	462 (2)	455 (1)	467 (2)	1.96
ANALYSIS	509 (5)	478 (2)	608 (6)	500 (4)	490 (3)	462 (1)	462 (1)	465 (1)	1.85
GEOMETRIE	508 (4)	478 (2)	608 (5)	508 (4)	488 (3)	464 (1)	462 (1)	478 (2)	1.92
STOCHASTIK	521 (6)	484 (3)	582 (7)	501 (5)	491 (4)	476 (2)	463 (1)	472 (2)	1.43
VERSCH	510 (5)	481 (2)	586 (6)	498 (4)	490 (3)	468 (1)	468 (1)	478 (2)	1.54

Maturandinnen und Maturanden des Schwerpunkts „Physik und Anwendungen der Mathematik“ haben in allen Bereichen des Mathematiktests klar am besten, die Gruppe „Bildnerisches Gestalten“ ebenfalls in allen Bereichen und teilweise zusammen mit der Gruppe „Philosophie/Pädagogik/Psychologie“ und dem Schwerpunkt „Musik“, am schlechtesten abgeschlossen (siehe Tabelle 14). Das Resultat der besten Gruppe fällt erwartungsgemäss aus. Die Effektstärke zwischen dieser und der schlechtesten Gruppe ist aber ausserordentlich gross. Die Schülerinnen und Schüler des Schwerpunkts „Alte Sprache“ haben am zweitbesten abgeschnitten, noch vor jenen des Schwerpunkts „Biologie und Chemie“ und jenen des Schwerpunkts „Wirtschaft und Recht“. Die Effektstärken der zweitbesten Gruppe „Alte Sprachen“ zur jeweils schlechtesten liegen immer noch bei erheblichen Werten zwischen d=0.71 (Stochastik) und d=0.55 (Verschiedenes).

Tab. 15: Vergleich der Testergebnisse in Biologie nach gewählten Schwerpunktfächern

Testbereiche	Alte Sprachen	Moderne Sprachen	Physik und Anwendungen der Mathematik	Biologie und Chemie	Wirtschaft und Recht	Philosophie/ Pädagogik/ Psychologie	Bildnerisches Gestalten	Musik	grösste Effektstärke
TOT_BIO	505 (4)	485 (2)	496 (3)	559 (5)	487 (2)	484 (2)	491 (2,3)	475 (1)	1.08
B_EVO	504 (3)	486 (1)	498 (2)	554 (4)	486 (1)	489 (1)	496 (2)	483 (1)	0.98
B_STRU	506 (3)	485 (1)	498 (2)	557 (4)	484 (1)	487 (1)	497 (2)	484 (1)	0.99
B_GEN	501 (3)	486 (1)	499 (3)	557 (4)	484 (1)	487 (1,2)	492 (2)	484 (1)	0.99
B_INFO	506 (5)	484 (1,2)	499 (4)	557 (6)	484 (1,2)	488 (2,3)	492 (3)	481 (1)	1.00
B_OEKO	506 (4)	484 (1)	498 (3)	555 (5)	484 (1)	488 (1,2)	493 (2,3)	484 (1)	0.95
B_STOFF	505 (4)	485 (1)	498 (3)	559 (5)	483 (1)	488 (1,2)	492 (2,3)	484 (1)	1.02

Im Biologietest haben die Maturandinnen und Maturanden des Schwerpunkts „Biologie und Chemie“ wiederum erwartungskonform in allen Bereichen die besten Ergebnisse erzielt (siehe Tabelle 15). Auch in diesem Test liegen die Schülerinnen und Schüler des Schwerpunkts „Alte Sprachen“ in allen Bereichen an zweiter Stelle, noch vor jenen des anderen naturwissenschaftlichen Schwerpunkts „Physik und Anwendungen der Mathematik“. Die Ergebnisse der Gruppe des Schwerpunkts „Musik“ sind im Gesamtergebnis klar, in den Teilbereichen teilweise zusammen mit den Gruppen „Moderne Sprachen“, „Philosophie/Pädagogik/Psychologie“ sowie „Wirtschaft und Recht“ am schlechtesten ausgefallen. Die Effektstärken der schlechtesten Gruppe zur „Spezialistengruppe“ sind sehr hoch und bleiben im Vergleich zur zweitbesten Gruppe „Alte Sprachen“ immer noch moderat (zwischen d=0.35 (Gesamttest) und d=0.22 (Genetik)).

Tab. 16: Vergleich der Testergebnisse im UFT nach gewählten Schwerpunktfächern

Testbereiche	Alte Sprachen	Moderne Sprachen	Physik und Anwendungen der Mathematik	Biologie und Chemie	Wirtschaft und Recht	Philosophie/ Pädagogik/ Psychologie	Bildnerisches Gestalten	Musik	grösste Effektstärke
U_QUANT	514 (4)	486 (2)	560 (5)	517 (4)	499 (3)	469 (1)	473 (1)	485 (2)	1.25
U_DIAGR	506 (4)	487 (3)	557 (6)	519 (5)	499 (4)	468 (1)	475 (1,2)	481 (2,3)	1.22
U_TEXT	514 (4)	488 (2)	553 (5)	519 (4)	496 (3)	476 (1)	478 (1)	482 (1,2)	1.02
U_NAT	511 (4)	488 (2)	549 (6)	521 (5)	496 (3)	474 (1)	477 (1)	485 (2)	1.01

Im UFT ist die Reihenfolge der Ergebnisse in allen Unterbereichen dieselbe (siehe Tabelle 16). Die Unterschiede zwischen der besten Gruppe („Physik und Anwendungen der Mathematik“) sowie der schlechtesten („Philosophie/Pädagogik/Psychologie“) sind erheblich (sehr grosse Effektstärke). Auch die Maturandinnen und Maturanden des Schwerpunkts „Bildnerisches Gestalten“ befinden sich in der Signifikanzgruppe mit den schlechtesten Ergebnissen.

Nachdem in den Vergleichen der Testergebnisse zwischen den Schwerpunktfach-Gruppen teilweise grosse Unterschiede ausgemacht werden können, interessiert, ob sich solche auch in den effektiv vergebenen Maturanoten in den drei Fächern zeigen (siehe Tabelle 17). Die Effektstärke zwischen den Extremgruppen bleibt beim Vergleich der Maturanote in Erstsprache eine mittlere. Hervorzuheben sind die vergleichsweise zur Testrangierung verbesserten Bewertungen für die Schülerinnen und Schüler des Schwerpunkts „Musik“ und das Abrutschen der Maturandinnen und Maturanden des Schwerpunkts „Biologie und Chemie“. Beim Vergleich der Maturanote in Mathematik halbiert sich der mit der Effektstärke gemessene Unterschied zwischen den beiden Extremgruppen gegenüber dem Testunterschied beinahe, er bleibt aber sehr hoch. In diesem Fach stimmt auch die sonstige Rangierung recht gut mit jener der Testergebnisse überein. Das deutet darauf hin, dass es den bewertenden

Personen in Mathematik eher gelingt, sich bei der Benotung implizit an objektiven Güte-Standards zu orientieren. Beim Vergleich der Matura-Erfahrungsnote in Biologie fällt auf, dass die Effektstärke beim Unterschied zwischen den Extremgruppen auf ein mittleres Mass sinkt. Die Rangierung stimmt mit kleinen Abweichungen erstaunlich gut mit der Rangierung des Biologietests überein. Die grösste Abweichung kann für die Schülerinnen und Schüler des Schwerpunkts „Biologie und Chemie“ sowie „Bildnerisches Gestalten“ festgestellt werden. Erstere wurden mittels Noten vergleichsweise schlechter bewertet (und diesbezüglich gar von den „Altsprachlern“ überholt), letztere wurden besser bewertet.

Tab. 17: Vergleich von Maturanoten zwischen den Gruppen verschiedener Schwerpunktfächer

Fächer	Alte Sprachen	Moderne Sprachen	Physik und Anwendungen der Mathematik	Biologie und Chemie	Wirtschaft und Recht	Philosophie/ Pädagogik/ Psychologie	Bildnerisches Gestalten	Musik	Effektstärke
Erstsprache	4.87 (5)	4.67 (3)	4.63 (2)	4.58 (1)	4.54 (1)	4.67 (3)	4.58 (1)	4.75 (4)	0.60
Mathematik	4.57 (4)	4.26 (2)	4.88 (5)	4.34 (3)	4.28 (2)	4.14 (1)	4.19 (1)	4.31 (2,3)	0.93
Biologie	4.91 (4)	4.61 (1)	4.69 (2)	4.80 (3)	4.65 (1,2)	4.66 (2)	4.81 (3)	4.61 (1)	0.54

Ergänzend zu diesen Mittelwertvergleichen sind aus Tabelle 18 die Prozentzahlen für ungenügende Ergebnisse ersichtlich. In Mathematik zeigen sich ungenügende Bewertungen in der schriftlichen Maturitätsprüfung für die Hälfte aller Maturandinnen oder mehr in den nach MAR 95 neuen Schwerpunkten „Philosophie/Pädagogik/Psychologie“, „Bildnerisches Gestalten“ und „Musik“. Im Weiteren haben offenbar Erfahrungsnoten kompensierende Wirkung für die in schriftlichen Prüfungen eher hohe Zahl an ungenügenden Bewertungen.

Tab. 18: Vergleich des Prozentanteils ungenügender Ergebnisse in den Maturaprüfungen zwischen den Gruppen unterschiedlicher Schwerpunktfächer

Schwerpunkte	Schriftliche Erstsprachprüfung	Gesamtmaturanote Erstsprache	Schriftliche Mathematikprüfung	Gesamtmaturanote Mathematik	Biologie-Erfahrungsnote	Gesamtmaturanote Naturwiss.
Alte Sprachen	9.7%	1.3%	32.3%	19.9%	6%	2.7%
Moderne Sprachen	19.2%	4.3%	46.7%	29.6%	8.7%	8.4%
Physik und Anwendungen der Mathematik	14.9%	4.8%	12.4%	6.3%	6.1%	2.3%
Biologie und Chemie	22.7%	6.2%	41.8%	17.2%	3.2%	3.5%
Wirtschaft und Recht	22.0%	5.8%	40.7%	25.5%	4.8%	5.2%
Philosophie/ Pädagogik/Psychologie	20.6%	3.7%	55.8%	32.3%	3.4%	8.9%
Bildnerisches Gestalten	24.2%	5.3%	48.8%	32.3%	2.5%	3.4%
Musik	13.0%	1.3%	48.6%	25.6%	5.4%	5.9%

4.3.3 Vergleiche nach Geschlecht

Wie bei den meisten Untersuchungen wurden die Testergebnisse und die Maturanoten auch auf Unterschiede zwischen den Geschlechtern untersucht. Über alle Vergleiche hinweg ergaben sich mehrheitlich bessere Resultate für die Maturandinnen in Erstsprache, hingegen mehrheitlich schlechtere in Mathematik, Biologie und im überfachlichen Fähigkeitstest. In Tabelle 19 sind nur jene Ergebnisse aufgeführt, bei denen die Effektstärke mindestens 0.2 beträgt. Biologie ist deshalb nicht dabei. Erwähnenswert ist trotzdem der Gegensatz zwischen dem signifikant schlechteren Abschneiden der Maturandinnen in allen Biologie-Testbereichen

und der signifikant besseren Erfahrungsnote in Biologie für diese Gruppe. In der Tabelle ebenfalls nicht aufgeführt ist die umgekehrt mit einem Effekt über .2 (d=.23) signifikant bessere Erfahrungsnote in Physik für die Maturanden und das Fehlen eines signifikanten Unterschiedes zwischen den Geschlechtern bei der Gesamtnote in Naturwissenschaften. Maturandinnen kompensierten offenbar die schlechteren Noten in Physik mit besseren Noten in Biologie.

Tab. 19: Signifikante Unterschiede zwischen den Geschlechtern mit Effektstärke >.2

Leistungs-ergebnisse	Geschlecht	Mittelwert	S	Mittlere Differenz	Effekt-stärke
Erstsprache Erf.-Note	männlich	4.49	.520	-.20	0.38
	weiblich	4.69	.498		
Erstsprache Maturanote	männlich	4.55	.547	-.14	0.25
	weiblich	4.69	.532		
TOT_MATH	männlich	523	96.26	49	0.56
	weiblich	474	81.76		
ANALYSIS	männlich	523	94.84	47	0.54
	weiblich	476	77.11		
GEOMETRIE	männlich	525	91.94	49	0.58
	weiblich	476	76.35		
STOCHASTIK	männlich	519	88.20	38	0.45
	weiblich	481	81.77		
M_VERSCH	männlich	515	86.13	35	0.42
	weiblich	480	78.20		
Math. schr. Maturaprüf.	männlich	4.19	1.052	.27	0.25
	weiblich	3.92	1.032		
U_QUANT	männlich	526	82.27	49	0.60
	weiblich	477	78.88		
U_DIAGR	männlich	525	79.56	48	0.61
	weiblich	477	78.12		
U_TEXT	männlich	522	82.10	42	0.52
	weiblich	480	78.18		
U_NAT	männlich	521	81.97	41	0.52
	weiblich	480	76.83		

4.3.4 Vergleiche nach Maturitätsquoten

Die Maturitätsquoten in der Schweiz unterscheiden sich beträchtlich. Im Maturajahrgang 2007 haben gemäss Bundesstatistik im Kanton Tessin 29.5% der jungen Erwachsenen einen gymnasialen Abschluss erworben, im Kanton Obwalden waren es lediglich 11.2%. Diese Unterschiede müssen nicht zwingend eine entsprechende Chancenungleichheit widerspiegeln. Es ist aber von Interesse, ob an den Maturaprüfungen überall die gleichen Anforderungen gestellt werden. Im Rahmen von EVAMAR II lassen sich Teile der Kompetenzen der Maturandinnen und Maturanden aus Kantonen mit unterschiedlichen Maturitätsquoten miteinander vergleichen. Weil die Stichprobe nicht für jeden Kanton repräsentativ ist, wird nur eine Grobeinteilung der Quoten zu Gruppen vorgenommen. Überlagerungen mit anderen Effekten bleiben dabei offen, ebenso die Frage der Gleichheit von Anforderungen bei der Eingangsselektion der Gymnasien. Bei der Beurteilung der Chancengleichheit müsste nämlich auch letztere berücksichtigt werden, weil ein „Aufholen“ von Kompetenzdefiziten während der Gymnasialzeit durchaus möglich ist, sodass sich dann Eingangsunterschiede am Schluss nicht mehr zeigen, die Chancenungleichheit aber bleibt. Wegen der sich überlagernden Effekte wird im Folgenden ein Vergleich zwischen Gruppen in einem oberen und einem

unteren Bereich vorgenommen. Orientiert man sich an einer möglichst gleichmässigen Dreiteilung, ergibt sich für das untere Drittel eine Quotengrenze von <17.5%, für das mittlere eine solche von 17.5%-18.9% und für das obere Drittel jene ab 19.0%.

Die Maturandinnen und Maturanden aus der Gruppe mit der tieferen Maturitätsquote erzielten in beinahe allen Test- und Notenbereichen die besseren Resultate. In Tabelle 20 sind jene Ergebnisse aufgeführt, bei denen die Effektstärke mindestens d=0.2 beträgt.

Tab. 20: Signifikante Unterschiede zwischen Kantonsgruppen mit unterschiedlichen Maturitätsquoten mit Effektstärke >.2

Leistungsergebnisse	Maturitätsquote	Mittelwert	S	Mittlere Differenz	Effektstärke
Erstsprache Erf.-Note	Unteres Drittel	4.67	.469	.16	0.31
	Oberes Drittel	4.51	.559		
Erstspr. schr. Matura-Prüf.	Unteres Drittel	4.50	.685	.20	0.28
	Oberes Drittel	4.30	.784		
Erstsprache Matura-Prüf.	Unteres Drittel	4.61	.614	.16	0.25
	Oberes Drittel	4.45	.666		
Erstsprache Matura-Note	Unteres Drittel	4.70	.499	.16	0.30
	Oberes Drittel	4.54	.572		
TOT_MATH	Unteres Drittel	504	89.73	25	0.26
	Oberes Drittel	479	92.71		
M_VERSCH	Unteres Drittel	507	83.72	30	0.35
	Oberes Drittel	477	82.92		
TOT_BIO	Unteres Drittel	517	76.85	50	0.61
	Oberes Drittel	467	86.02		
B_EVO	Unteres Drittel	515	70.63	44	0.58
	Oberes Drittel	471	81.35		
B_STRU	Unteres Drittel	516	72.77	46	0.59
	Oberes Drittel	470	83.12		
B_GEN	Unteres Drittel	515	73.03	43	0.55
	Oberes Drittel	472	83.32		
B_INFO	Unteres Drittel	516	72.78	47	0.60
	Oberes Drittel	469	84.35		
B_OEKO	Unteres Drittel	515	73.45	46	0.57
	Oberes Drittel	469	85.58		
B_STOFF	Unteres Drittel	516	73.49	47	0.59
	Oberes Drittel	469	86.04		
U_DIAGR	Unteres Drittel	503	81.78	17	0.22
	Oberes Drittel	486	75.28		
U_TEXT	Unteres Drittel	504	81.65	21	0.26
	Oberes Drittel	483	76.27		
U_NAT	Unteres Drittel	504	80.39	19	0.25
	Oberes Drittel	485	74.35		

Die grössten Unterschiede lassen sich bei den Ergebnissen in Biologie finden (mindestens mittlerer Effekt). Bei den meisten Maturanoten fallen die Unterschiede in der gleichen Richtung aus wie bei den Testergebnissen, aber in geringerem Ausmass. Ausnahme davon sind die Erstsprachennoten, deren Differenzen die Unterschiede in den Testergebnissen in Erstsprache sogar übersteigen. Es kann somit festgehalten werden, dass höhere Maturitäts-

quoten mehrheitlich mit schlechteren Ergebnissen der in EVAMAR II erfassten Teile der Maturitätsausbildung verknüpft sind und sich damit die Frage der Chancengleichheit stellt.

4.3.5 Vergleiche nach Strukturelement Lang- oder Kurzzeitgymnasium

Ein Teil der Schweizer Gymnasien schliesst direkt an die Primarstufe an und integriert die Sekundarstufe I in ein Langgymnasium. Es drängte sich deshalb auf zu untersuchen, ob sich Unterschiede in den Resultaten zu den Kurzgymnasien ergeben. Der Vergleich zeigt, dass in fast allen Testbereichen und auch bei den Maturanoten die Langgymnasien zwar signifikant besser abschneiden, die Effektstärken aber sehr gering sind und nur in den in Tabelle 21 aufgeführten Bereichen das Mass von .2 (geringer Effekt) überschreiten. Sie finden sich durchwegs im Biologietest (nicht aber in den Biologienoten) und im Teil „Grammatik“ des Erstsprachtests. Vermutlich wird im letzten Bereich am Langgymnasium ein solideres Fundament gelegt als in der Sekundarschule. Ausserdem findet möglicherweise in Biologie am ehesten ein systematischer, über die ganze Dauer des Gymnasiums geplanter Aufbau statt, der zu einer breiteren Wissensbasis in diesem Fach führt. Interessanterweise ist aber auf der anderen Seite die Erfahrungsnote in Biologie die einzige Note, die für die Kurzgymnasien vorteilhafter ausfällt (d=.17). Vermutlich zeigt sich in diesem widersprüchlichen Ergebnis am deutlichsten die teilweise Anpassung der Notengebung an die Leistungsfähigkeit der Klasse, wenn klare Anforderungsstandards fehlen.

Tab. 21: Signifikante Unterschiede zwischen Lang- und Kurzzeitgymnasien mit Effektstärke >.2

Leistungs-ergebnisse	Dauer	Mittelwert	Standard-abweichung	Mittlere Differenz	Effektstärke
GRAM	Kurzzeit_G	498	73.11	-16	0.22
	Langzeit_G	514	78.03		
TOT_BIO	Kurzzeit_G	493	84.51	-27	0.32
	Langzeit_G	520	80.18		
B_EVO	Kurzzeit_G	494	78.86	-24	0.31
	Langzeit_G	518	76.34		
B_STRU	Kurzzeit_G	494	80.80	-24	0.30
	Langzeit_G	518	77.19		
B_GEN	Kurzzeit_G	494	80.88	-24	0.30
	Langzeit_G	518	77.21		
B_INFO	Kurzzeit_G	493	81.64	-25	0.31
	Langzeit_G	518	76.15		
B_OEKO	Kurzzeit_G	493	81.80	-25	0.31
	Langzeit_G	518	77.55		
B_STOFF	Kurzzeit_G	493	82.96	-25	0.31
	Langzeit_G	518	76.90		

5 Ergebnisse III: Analyse der Maturaprüfungen – TP D1

Im Projektteil D1 von EVAMAR II wurden die schriftlichen Maturaprüfungen der gleichen Stichprobe wie in Teilprojekt C (inklusive Tessin) qualitativ untersucht im Hinblick darauf, welche fachinhaltlichen und welche kognitiven Anforderungen sie stellen, und inwieweit es eine Übereinstimmung gibt mit den in den Teilprojekten A und B ermittelten Komponenten der Studierfähigkeit sowie mit den in Teilprojekt C durchgeführten Tests. Diese Analyse erfolgte zunächst für die Testfächer Erstsprache, Mathematik und Biologie; sie könnte in einem Folgeprojekt auf weitere Fächer ausgedehnt werden.

Die schriftlichen Schweizer Maturaprüfungen in Erstsprache zeigen sich in ausserordentlicher Breite und Vielfalt: Es scheint, als ob jede Schule und jeder Kanton, ja jede Lehrperson eigene Vorstellungen darüber hat, wie eine schriftliche Maturaprüfung auszugestalten sei und unter welchen Bedingungen die sogenannte „Hochschulreife" überprüft werden soll. Die grösste Übereinstimmung hat die Schweizer Lehrerschaft darin, dass sie hauptsächlich Texte in der jeweiligen Originalsprache und selten Übersetzungen vorlegt, sowie darin, dass sie selten Texte mit weiblicher Autorschaft berücksichtigt. Ziemlich einig sind sich die Schweizer Lehrpersonen auch darin, dass sie meistens einen bis zwei Primärtexte aus dem 20. oder 21. Jahrhundert zu einem möglichst aktuellen oder zeitlosen Thema vorlegen. Dabei wählen sie zu rund 40-60% literarische Primärtexte aus und nur zu 10-20% Sachtexte; den Rest bilden sogenannte „Philosophische Texte" (Deutsch 12%, Französisch 14%, Italienisch 30%). Vom Anspruchsniveau her gesehen, befinden sich die meisten der schriftlichen Maturaprüfungen (D 77%, F 99.1%, I 77%) im oberen Bereich der kognitiven Anforderungen (Niveau 4-6 nach Anderson u. a.). Die grössten regionalen Unterschiede liegen in der eindeutigen Dominanz des Anforderungsbereiches „Erschaffen-Bewerten" in der Romandie (zwei Drittel), im Unterschied zum dritten Rang dieses Bereiches im Tessin (ein Fünftel). In der Deutschschweiz ist die Verteilung ausgeglichener, hier werden zu einem Drittel Beurteilungsleistungen gefordert, zu einem Viertel Analysen und zu einem weiteren Viertel das Erfinden von Produkten. Insgesamt stellt die vierstündige Aufsatzform ein Produkt der höchsten kognitiven Stufe, jener des Erschaffens, dar. Inhaltlich zeugen die vorgelegten Themen von einer grossen Aktualität, der Orientierung an der Lebenswelt der Jugendlichen, einer starken Betonung der Meinungsvielfalt und von einer intensiven Auseinandersetzung sowohl mit wichtigen gesellschaftlichen Fragen als auch mit zeitlosen Sinnfragen des Menschseins. Im Vergleich mit den Ergebnissen aus der Dozierendenbefragung stehen allerdings die ermittelten vielfältigen Prüfungsinhalte (Autorinnen und Autoren, Gesellschaftsthemen, ethische Probleme u. a.) im Grossen und Ganzen nur beschränkt im Erwartungshorizont der befragten Dozierenden aller Fachrichtungen. Das Leseverständnis wird zudem vor allem an literarischen und nur wenig an fachwissenschaftlichen Texten geprüft.

40.8% der untersuchten Aufgaben in Schweizer Biologie-Maturaprüfungen enthalten „higher-order questions" (Hauptstufen 3-6 bei Anderson u. a.). Dieser Wert macht sie vergleichbar mit den bekannten, gross angelegten US-amerikanischen Assessments wie beispielsweise dem Medical College Admission Test (MCAT) oder dem „Advanced Placement (AP) Biology course. Und das ist aus zweierlei Hinsicht nicht selbstverständlich:

a) Deutlich später als im amerikanischen Kontext wurden erst in den 80er Jahren im deutschsprachigen Raum Ansätze zur Förderung des produktiven Denkens bzw. so genannte Denktrainings entwickelt. Erfahrungsgemäss kann es Jahrzehnte dauern, bis sich selbst empirisch gut abgesicherte universitäre Überlegungen in der tatsächlichen Lehrerbildung niederschlagen. Gemäss den vorliegenden Ergebnissen scheint der Transfer in die schweizerische SekII-Didaktikausbildung aber grossflächig abgeschlossen zu sein und in vielen gymnasialen Schulzimmern Früchte zu tragen.
b) Das Verfassen und die Korrektur von guten Transferaufgaben ist heikel und sehr zeitintensiv. Offenbar sind viele Lehrpersonen davon überzeugt, dass sich der Aufwand lohnt.

Der hohe Prozentwert widerlegt auch das altbekannte Klischee, wonach in Biologie durch blindes Auswendiglernen leicht eine gute Note zu erreichen und „Denken" unnötig sei. Die Resultate der Analyse der Biologieprüfungen weisen aber auch auf eine möglicherweise ungenügende Berücksichtigung der Themenbereiche „Ökologie und Systematik/Evolution" hin.

Auch bei den Mathematikprüfungen liegt inhaltlich eine grosse Vielfalt vor. Die Maturaaufgaben zur Mathematik sind in aller Regel sorgfältig konstruiert und sehr phantasievoll ausgearbeitet. Es gibt daneben aber auch eine stattliche Zahl von „Standardprüfungen": je eine Aufgabe zur Kurvendiskussion, zur Vektor- und zur Wahrscheinlichkeitsrechnung, zur

Differentialrechnung (Extremalproblem) sowie eine Aufgabe mit mehreren Kurzaufgaben, die weitere Gebiete (Folgen und Reihen, elementare Algebra etc.) abdeckt. Bezüglich des kognitiven Anspruchsniveaus ist bei diesen Prüfungen der Anteil an „Erinnern" besonders hoch. Grosse Unterschiede gibt es bei den (potentiellen) kognitiven Anforderungen sowohl zwischen Grundlagenfachprüfungen und Schwerpunktfachprüfungen als auch zwischen Grundlagenfachprüfungen verschiedener Schulen. Hinsichtlich des kognitiven Anspruchspotentials ist der Grossteil der Matura(teil-)aufgaben der Kategorie „Ausführen; eine Prozedur auf ein bereits vertrautes Problem anwenden" zuzuordnen. Die fast in jeder Prüfung vorkommende Extremalaufgabe gehört unserer Einschätzung nach meist in die Kategorie „Ein neues Problem auf ein bereits bekanntes Problem zurückführen". Bei anspruchsvollen Extremalproblemen ist zudem ein hohes Mass an Modellierungsarbeit zu leisten, was dem Niveau „Modellieren, bekannte mathematische Modelle einsetzen" (falls das Modell bereits bekannt ist) oder aber dem Niveau „Modellieren, neue mathematische Modelle einsetzen" entspricht. Solche Aufgaben kommen fast nur in Prüfungen im Schwerpunktfach „Physik und Anwendungen der Mathematik" vor.

6 Ergebnisse IV: Analyse der Maturaarbeiten – TP D2

Dieses Teilprojekt wurde vollständig durch die Pädagogische Hochschule der Fachhochschule Nordwestschweiz konzipiert und durchgeführt.

Die Analyse der Qualität der Maturaarbeiten aus der Deutschschweiz, der Romandie und dem Tessin basierte auf drei Elementen: Im Zentrum der Untersuchung stand die Analyse von 437 Maturaarbeiten, die jeweils von zwei unabhängigen Expertinnen oder Experten einem Rating unterzogen wurden. Die Verfasserinnen und Verfasser dieser Arbeiten waren zudem Teil einer Stichprobe von Maturandinnen und Maturanden, die schriftlich zum Kontext, zum Verlauf und zu den Resultaten ihrer Projekte befragt wurde. Ergänzt werden diese Befunde durch die Rekonstruktion institutioneller Rahmenbedingungen auf nationaler und kantonaler Ebene sowie auf der Ebene von 30 an der Untersuchung beteiligten Gymnasien.

Die Evaluation der Maturaarbeiten kommt insgesamt zu einem positiven Befund: Es zeigt sich, dass die grosse Mehrheit der Maturandinnen und Maturanden mit den Grundregeln wissenschaftlichen Arbeitens vertraut ist. Die Analyse erhärtet die Hypothese, dass die inhaltliche Qualität der Arbeiten mit ihrer formalen und ihrer sprachlichen Qualität korreliert: Schülerinnen und Schüler, die bei den inhaltlichen Gütekriterien gut abschneiden, erzielen in der Regel auch bei den anderen Gütekriterien gute Werte, hingegen sind bei inhaltlich schwachen Arbeiten auch überdurchschnittlich viele formale und sprachliche Mängel festzustellen. Es wirkt sich positiv auf die Qualität der Maturaarbeiten aus, wenn die Schulen die Themenwahl durch die Vorgabe von Rahmenthemen strukturieren; fehlt eine solche institutionelle Beschränkung, so fallen die Arbeiten im Durchschnitt etwas schlechter aus. Dies dürfte nicht zuletzt darauf zurückzuführen sein, dass mit der Themenbreite auch die Komplexität der Betreuungsaufgaben zunimmt.

Zwischen natur-, sozial- und geisteswissenschaftlich ausgerichteten Arbeiten sind keine gravierenden qualitativen Differenzen festzustellen. Auch beim Geschlechtervergleich ergeben sich keine bedeutsamen Qualitätsunterschiede. Anders verhält es sich beim Regionalvergleich: Wiewohl der Befund insgesamt für alle Landesteile positiv ausfällt, stechen doch beträchtliche sprachregionale Unterschiede zwischen der Deutschschweiz und der lateinischen Schweiz ins Auge, die statistisch im Rahmen dieses Projekts nicht vollends aufgeklärt werden konnten. Es ist zu vermuten, dass das bessere Abschneiden der Romandie und des Kantons Tessin durch engere Vorgaben bei der Themenwahl und durch die Einbettung der Arbeit in seminarähnliche Strukturen begünstigt wird. Es kann jedoch nicht aus-

geschlossen werden, dass die Ergebnisse des Ratings durch unterschiedliche sprachregionale Bewertungstraditionen beeinflusst worden sind.

In Hinblick auf die Rahmenbedingungen ist festzustellen, dass die Maturandinnen und Maturanden schulische Leitfäden und andere offizielle Papiere eher beiläufig zur Kenntnis nehmen; Transparenz und Verhaltenssicherheit ergibt sich für sie am ehesten auf der Basis interpersonalen Kontakts mit den betreuenden Lehrpersonen. Dabei ist weniger die Art und Frequenz der Kontakte entscheidend, sondern vielmehr die wahrgenommene Intensität der Unterstützung bei der Bewältigung kritischer Phasen. Dies beginnt bei der Formulierung des Themas und erstreckt sich über die Klärung von Methodenfragen bis hin zum formalen Aufbau des Berichts. Ähnlich wie die Frequenz der Kontakte mit Lehrpersonen ist auch die Bearbeitungsdauer der Projekte für die Qualität der Maturaarbeiten nur wenig bedeutsam.

In der Analyse fällt auf, dass das Rating der Expertinnen und Experten in der Regel kritischer ausfällt als die Benotung durch die Lehrpersonen. Dabei ist zu berücksichtigen, dass der Fokus der pädagogischen Bewertung sich vom Fokus der vorliegenden Untersuchung in mehrfacher Hinsicht unterscheidet. So befinden die Lehrpersonen in der Regel nicht allein über das schriftliche Produkt, sondern auch über die mündliche Präsentation. Zudem musste sich die vorliegende Qualitätsanalyse auf eine summative Dimension beschränken, da sie nicht prozessbegleitend angelegt werden konnte, während die Lehrpersonen auch den Lernfortschritt der Schülerinnen und Schüler im Projektverlauf vor Augen haben. Darüber hinaus ist zu betonen, dass die im Rahmen dieser Untersuchung durchgeführte Evaluation der Maturaarbeiten sich weitgehend auf den akademischen Blickwinkel des Erwerbs der Studierfähigkeit beschränkt, während die Lehrpersonen bei der Wahrnehmung ihres Bildungsauftrags eine Vielzahl anderer Aspekte in die Bewertung mit einfliessen lassen können.

Die Schülerinnen und Schüler tendieren dahin, der Maturaarbeit sowohl einen intrinsischen wie auch einen extrinsischen Wert zuzusprechen; nur in vergleichsweise wenigen Fällen wird das Projekt im Rückblick als nutzlos erachtet. Die inhaltliche Qualität korreliert positiv mit den Einschätzungen der Schülerinnen und Schüler zum Nutzen der Arbeiten für die weitere akademische oder berufliche Qualifikation wie auch mit dem Nutzen für die persönliche Entwicklung.

Von der Warte tertiärer Bildungsinstitutionen aus ergibt sich für die Maturaarbeit prima facie das Bild, dass die Maturandinnen und Maturanden in der Lage sein sollten, den elementaren Anforderungen eines Hochschulstudiums gerecht werden zu können: Sowohl bei der Strukturierung von Texten wie auch bei der Verwendung von Zitaten und beim Anlegen von Bibliographien legen die Maturaarbeiten davon Zeugnis ab, dass die gängigen akademischen Regeln bereits im Gymnasium eingeübt werden. Dasselbe gilt für die Wiedergabe elementarer theoretischer und empirischer Sachverhalte sowie ansatzweise auch für die Gewinnung von Informationen durch Recherchen oder durch Experimente, wiewohl bei den schulischen Projekten deskriptive und reproduktive Anteile deutlich gewichtiger sind als analytische Anteile.

Indes ist hier zu betonen, dass eine Querschnittsanalyse zu den Maturaarbeiten keine Prognose im Hinblick auf die Nachhaltigkeit der Lernprozesse erlaubt. Die Reichweite der institutionellen Vorkehrungen, die Intensität der pädagogischen Betreuung und die Unterstützung durch das soziale Umfeld bilden Kontextfaktoren, deren Zusammenwirken eine hohe Qualität des Outcome begünstigt. Wirken diese Faktoren im Rahmen des Studienalltags nicht mehr gebündelt zusammen, ist nicht auszuschliessen, dass viele der ersten Studienarbeiten von Anfangssemestern hinter dem Niveau der Maturaarbeiten zurückbleiben.

7 Grenzen der Studie EVAMAR II

An dieser Stelle soll nochmals ausdrücklich auf die schon im zweiten Abschnitt erwähnten Grenzen der Studie EVAMAR II hingewiesen werden. Grundsätzlich können nur über einige ausgewählte Aspekte der Ziele und der Zielerreichung der gymnasialen Bildung Aussagen gemacht werden. Insbesondere folgende Komponenten einer umfassenden Studierfähigkeit wurden im Projekt EVAMAR II nicht untersucht und ihr Vorhandensein bei Maturandinnen und Maturanden nicht analysiert: erstens Wissen und Können, welches zwar Teil einer breiten Allgemeinbildung ist – als Grundlage zur Lösung anspruchsvoller Aufgaben in der Gesellschaft (vertiefte Gesellschaftsreife) – und damit in der gymnasialen Bildung vermittelt werden muss, aber an der Universität nicht direkt vorausgesetzt wird; zweitens zur Studierfähigkeit gehörendes notwendiges Wissen und Können aus anderen Unterrichtsfächern als Erstsprache, Mathematik und Biologie; drittens Elemente überfachlicher Kompetenzen, die im gegebenen Projektrahmen nur schlecht erfassbar gewesen wären, dennoch aber für die erfolgreiche Bewältigung eines Studiums wichtig sind (z. B. die Fähigkeit des selbstorganisierten Lernens, der allgemeinen Selbstorganisation, des Recherchierens, der raschen Verarbeitung umfangreicher Fachliteratur usw.). Die Bedeutung solcher Faktoren liesse sich nur durch weitere, im umfassenden Konzept vorgeschlagene Untersuchungen ermitteln (Längsschnittuntersuchungen bei Studierenden), und sie wären nur durch aufwändige Assessmentverfahren zu messen. Immerhin wurden solche Faktoren aber bereits im Projekt EVAMAR I über Befragungen, also über Selbsteinschätzungen und damit annäherungsweise valide ermittelt. Im Zusatzfragebogen von EVAMAR II wurden zudem diesbezüglich ergänzende Fragen zu EVAMAR I eingestreut.

8 Die Ergebnisse im Überblick: Zusammenfassung

Grundlagen und Voranalysen

Die Test- und Befragungserhebungen bei einer Stichprobe von rund 3’800 Maturandinnen und Maturanden aus der Deutschschweiz und der Romandie und die Analysen von schriftlichen Maturaprüfungen aller drei Sprachregionen stützen sich auf folgende Grundlagen: Analyse von Lehrmaterialien und ersten Zwischenprüfungen der 16 gemessen an den Studierendenzahlen grössten universitären Studienfächer der Schweiz, Befragung der Dozierenden der Lehrveranstaltungen dieser Fächer und Einbezug der aktuellen Theorie und Forschung zur Messung kognitiver Leistungen und zur Voraussage des Studienerfolgs.

Das Hauptgewicht des Tests in Erstsprache liegt auf übergreifenden, für alle Studienrichtungen wesentlichen Sprachkompetenzen.

Der Mathematiktest enthält vor allem Aufgaben zu Inhalten, die für eine breite Anzahl von Studienfächern und nicht speziell für das Studienfach Mathematik von Bedeutung sind.

Der Biologietest hat den Charakter eines Wissenstests zu Fachinhalten, die vor allem für ein Biologie- oder Medizinstudium, aber auch teilweise für nicht naturwissenschaftliche Studien von Bedeutung sind.

Der überfachliche Fähigkeitstest (UFT) lehnt sich an die Eignungsprüfung für das Medizinstudium an und misst allgemeinere kognitive Fähigkeiten. Er wurde nur ergänzend in der Deutschschweiz durchgeführt.

Das Testinstrumentarium umfasst somit verschiedenartige Prädiktoren des Studienerfolgs und kann deshalb als ausgewogen bezeichnet werden. Interpretiert man allgemeine Studierfähig-

keit dahin gehend, jedes beliebige Studium ohne Probleme aufnehmen zu können, haben die in den Testbereichen gemessenen Kompetenzen folgende Bedeutung: Erstsprachkompetenzen gehören zu den Grundlagen beinahe jedes Studienfachs. Mathematik ist wichtig für eine grosse Zahl von Fächern, in denen die mathematische Formalsprache verwendet wird. Biologisches Wissen braucht es nur für eine eingeschränkte Zahl von Studienfächern. Die mit dem UFT gemessenen Fähigkeiten korrelieren nachgewiesenermassen signifikant mit dem Studienerfolg im Medizinstudium und damit mit jenem in allen anderen Studienrichtungen, die gleiche oder ähnliche Fähigkeiten erfordern. Die Voranalysen haben im Weiteren ergeben, dass es allgemein für jedes Fachstudium von Vorteil ist, bereits zu Beginn über Wissen und Können aus dem entsprechenden Maturitätsfach – wenn auch in unterschiedlichem Umfang − zu verfügen.

Gesamtergebnisse der Tests

Im Mittel wurden die Testaufgaben etwa zur Hälfte richtig gelöst; die Resultate waren in Erstsprache etwas besser als in Mathematik und Biologie. In Biologie sind die Ergebnisse für die Gruppe der Maturandinnen und Maturanden, die dieses Fach bereits vor einem halben Jahr oder mehr abgeschlossen hatten, klar schlechter ausgefallen, vermutlich in Folge des Vergessenseffekts. Das deutet darauf hin, dass das an Maturitätsschulen erworbene Wissen nur für kurze Zeit vollumfänglich präsent ist, und dass sowohl ein möglichst später Abschluss des Faches am Gymnasium wie auch eine rasche Aufnahme des Studiums vorteilhaft sind. Die Ergebnisse des UFT liegen nur leicht unter den Resultaten des jährlich durchgeführten Medizineignungstests. Das ist plausibel, weil bei den EVAMAR-Erhebungen auch alle jene Personen teilgenommen haben, die sich ein Medizinstudium wegen mangelnden Fähigkeiten nicht zutrauen. Dieses Ergebnis ist zudem ein wichtiger Indikator für die hohe Anstrengungsbereitschaft, welche die Maturandinnen und Maturanden bei der Bearbeitung der EVAMAR-Tests offenbar an den Tag legten. Aufgrund der Nähe der Ergebnisse des UFT zu den Resultaten, wie sie im individualbiographisch hochrelevanten Kontext der Eignungsprüfung erreicht werden, braucht diese den Vergleich mit jener in den gewohnten notenwirksamen Testsituationen im Gymnasium keineswegs zu scheuen. Die Schweizer Maturandinnen und Maturanden erzielten somit in den untersuchten Bereichen Ergebnisse, die für standardisierte, für eine bestimmte Population entwickelte Tests normal sind. Zu dieser Normalität gehört, dass die Aufgaben im Mittel etwa zur Hälfte richtig gelöst wurden. Insofern kann der Ausbildungsstand als zufriedenstellend bewertet werden.

Sowohl zwischen Einzelpersonen als auch zwischen ganzen Klassen gibt es grosse Unterschiede in den Ergebnissen. Diese fallen grösser aus als bei den effektiv erteilten Maturanoten. Es lässt sich somit feststellen, dass erhebliche Anteile von Maturandinnen und Maturanden in den mit den Tests erfassten Bereichen über vermutlich ungenügende Kompetenzen in mindestens einem Testbereich verfügen. Diese Erkenntnis ergibt sich aber auch schon bei einer Analyse der durch die Schulen selbst vergebenen Maturanoten, am ausgeprägtesten in Mathematik. 41.4% der Maturandinnen und Maturanden wurden im Jahre 2007 in der schriftlichen Maturaprüfung mit einer ungenügenden Note bewertet. Bei der Maturaendnote für Mathematik, welche auch die Erfahrungsnoten und die mündliche Prüfung berücksichtigt, waren es immer noch 24.4%. Es lässt sich somit feststellen, dass nicht alle Maturandinnen und Maturanden in der ganzen Breite über genügende Kompetenzen verfügen, um jedes beliebige Studium aufnehmen zu können, ohne zusätzlich Lücken füllen zu müssen; dies obwohl sie durch die Verleihung des Maturaausweises als „allgemein studierfähig“ qualifiziert wurden. Diese Erkenntnis stimmt überein mit von Dozierenden subjektiv festgestellten Kompetenzlücken in den Bereichen Mathematik und Erstsprache. Wegen des Kompensationssystems für ungenügende Noten an Maturaprüfungen ist dieses Ergebnis der Maturitätsbildung aber kaum vermeidbar.

Ergebnisse der Tests unter strukturellen Gesichtspunkten

Die Anzahl Jahre der gymnasialen Schuldauer an einem eigenen gymnasialen Schulort scheint sich auf die Testergebnisse auszuwirken. Eine Dauer von mindestens vier Jahren an einem Gymnasium geht einher mit besseren Ergebnissen im Vergleich zu jener Ausbildungsstruktur, bei der das erste von vier Jahren in einer „gymnasialen" Klasse an der örtlichen Sekundarschule absolviert werden kann. Dies zeigt sich klar beim Mathematiktest, in leichterer Ausprägung auch bei Erstsprache und in der Deutschschweiz zusätzlich beim UFT.

Langzeitgymnasien schneiden in fast allen Testbereichen besser ab als die Kurzzeitgymnasien, allerdings nur leicht. Einzig im Biologietest ergeben sich auch grössere bzw. beinahe mittlere Differenzen zugunsten der Langzeitgymnasien, obwohl bei Letzteren die Biologie-Erfahrungsnoten im Ausmass zwar leicht, aber signifikant schlechter ausgefallen sind. Dabei dürfte es sich aber eher um eine „Notenanomalie" bzw. das Ergebnis einer klassenorientierten Benotung handeln.

Zwischen den verschiedenen Schwerpunktfachgruppen lassen sich z. T. erhebliche Unterschiede feststellen. Die Spitzenergebnisse in Mathematik der Gruppe des Schwerpunktfachs (SPF) „Physik und Anwendungen der Mathematik" (PAM) und in Biologie der Gruppe des SPF „Biologie und Chemie" (BCH) waren dabei systembedingt zu erwarten. Der Spitzenwert für die Maturandinnen und Maturanden des SPF „Alte Sprachen" im Vergleich zu jenen des SPF „Moderne Sprachen", die beim Erstsprachtest lediglich eine mittlere Rangierung erreichten, überrascht jedoch teilweise. Auch für weitere Ungleichheiten in einigen Kompetenzbereichen gibt es keine Systemerklärungen. Die Gruppe des SPF „Alte Sprachen" hat unter dem Aspekt der Ausgeglichenheit am besten abgeschnitten, die Maturandinnen und Maturanden der neuen Schwerpunktfächer „Musik", „Bildnerisches Gestalten", „Philosophie/Pädagogik/Psychologie" (PPP) eher schlechter. Bei der Analyse der effektiv erteilten Maturanoten ergeben sich im Vergleich zu den Testergebnissen einige Rangverschiebungen, und vor allem erweisen sich die Unterschiede als geringer ausgeprägt. Es fällt aber auch auf, dass es SPF-Gruppen mit beinahe 50% („Moderne Sprachen", „Bildnerisches Gestalten", „Musik") oder gar über 50% (PPP) ungenügenden Noten in der schriftlichen Mathematikprüfung gibt. Bei der Maturagesamtnote in Mathematik sind es für diese SPF-Gruppen (jetzt ohne „Musik") immer noch rund 30% ungenügende Noten. Mit Abstand die beste Gruppe beim UFT war jene des SPF PAM. Die Gruppen der neuen Schwerpunktfächer PPP und „Bildnerisches Gestalten" sowie in einem Bereich auch „Musik" rangieren dagegen am Ende.

Die Unterschiede zwischen den Geschlechtern stimmen mit bisherigen Untersuchungen gut überein. Maturandinnen haben eher in Erstsprache, Maturanden klar in Mathematik und im naturwissenschaftlich ausgerichteten UFT besser abgeschnitten.

Zieht man einen Vergleich zwischen den Kantonen, die sich hinsichtlich der Maturitätsquoten im oberen Drittel befinden, und jenen im unteren Drittel, so zeigen sich leichte bis zwischen mittel und gross liegende Unterschiede. Die positiven Differenzen zugunsten der Kantone mit den niedrigsten Maturitätsquoten stimmen nur teilweise mit entsprechenden Unterschieden bei den Maturanoten überein, welche aber grossmehrheitlich bei dieser Gruppe im Mittel ebenfalls besser ausfallen. Es stellt sich die Frage der Chancengleichheit beim Erwerb der Berechtigung zum Hochschulzugang.

Notengebung

Die Notengebung stimmt nur teilweise mit den Testergebnissen überein. Eine erste Erklärung dafür ist jene, dass mit den Tests teilweise andere Kompetenzen gemessen wurden als durch die Maturaprüfungen. Diese Unterschiede sind bedingt durch den eingeschränkten Fokus der Untersuchung und haben nichts mit mangelnder Zielerreichung der Gymnasien zu tun. Eine zweite Deutung der Befunde setzt beim Verfahren der Notengebung an. Einerseits kann eine teilweise Anpassung der Notengebung an die Leistungsfähigkeit der Klasse (Sozialnorm-orientierung) vermutet werden, was im Vergleich zur Orientierung an Klassen übergreifenden Gütestandards (Kriteriumsorientierung) zu verschiedenen Notenmassstäben führt. Dennoch lassen sich andererseits auch bei den Notenvergleichen strukturelle Übereinstimmungen mit den Testergebnissen feststellen, was annehmen lässt, dass sich die Lehrerinnen und Lehrer bei der Notengebung zu einem beträchtlichen Teil auch an „objektivierten Gütestandards" orientieren. Im Weiteren sind bei der Notengebung unerklärliche „Anomalien" zu finden, welche wiederum eine nur teilweise Kriteriumsorientierung annehmen lassen. Dazu gehört z. B. die Erfahrungsnote in Biologie, welche in den Schwerpunktgruppen „Alte Sprachen" und „Bildnerisches Gestalten" höher liegt als in der Gruppe „Biologie und Chemie". Auffällig ist schliesslich, dass schriftliche Maturaprüfungen vor allem in Mathematik, aber auch in Erstsprache zu einem hohen Anteil an ungenügenden Ergebnissen führen, diese aber jeweils durch die Note der mündlichen Prüfungen und die Erfahrungsnoten „aufgebessert" werden.

Schriftliche Maturaprüfungen

Die schriftlichen Maturaprüfungen sind in allen untersuchten Fächern auf der einen Seite häufig anspruchsvoll, erfordern zur Lösung verschiedene kognitive Fähigkeiten (eine reine Wiedergabe von zuvor auswendig gelerntem Wissen ist nicht ausreichend) und decken Bereiche ab, die für die Studierfähigkeit von Bedeutung sind. Auf der anderen Seite konnten viele Prüfungen gefunden werden, welche diesem Bild nicht entsprechen. Insgesamt sind die Aufgabenstellungen recht heterogen.

Maturaarbeiten

Die wissenschaftspropädeutische Qualität der untersuchten Maturaarbeiten ist mehrheitlich als zufriedenstellend einzustufen. Obwohl der Beweis der Nachhaltigkeit bislang fehlt, sprechen viele Indizien dafür, dass es sich bei der Maturaarbeit um eine im Hinblick auf die Studierfähigkeit sinnvolle und ertragreiche Lern- und Prüfungsform handelt.