Institut zur Qualitätsentwicklung
im Bildungswesen

Gemeinsame Abituraufgabenpools der Länder

Evaluation von Aufgaben der Pools für das Prüfungsjahr 2017

Ergebnisse zur Bewährung der Aufgaben

Dr. Lars Hoffmann, Dr. Pauline Schröter, Prof. Dr. Petra Stanat

Inhalt

1 Kurzzusammenfassung

Der vorliegende Bericht stellt die abschließenden Ergebnisse der Evaluation dar, in der untersucht wurde, wie sich die Aufgaben der Abituraufgabenpools für das Prüfungsjahr 2017 bewährt haben. Zumeist sind in Bezug auf die in sechs Evaluationsbereichen betrachteten Aspekte nur geringe Unterschiede zwischen Aufgaben aus den Pools und landeseigenen Aufgaben festzustellen, sodass in dieser Hinsicht insgesamt ein positives Fazit zum erstmaligen Einsatz von Aufgaben aus den Pools in den Abiturprüfungen der Länder gezogen werden kann. Allerdings sind fachspezifische Auffälligkeiten zu verzeichnen, die wichtige Hinweise für den weiteren Prozess der gemeinsamen Entwicklung von Abituraufgaben geben:

- In den Abiturprüfungen der Länder wurden im Fach Deutsch aus dem Pool fast ausschließlich Aufgaben zum materialgestützten Schreiben und textbezogene Aufgaben zu nicht-literarischen Texten eingesetzt. Da diese Aufgaben (im Vergleich zu den Aufgaben zu literarischen Texten) häufiger von Schülerinnen und Schülern mit weniger guten Vorleistungen gewählt wurden, sind bei den Poolaufgaben insgesamt im Mittel weniger Notenpunkte erzielt worden als bei den landeseigenen Aufgaben.
- Von den Lehrkräften wurden die Poolaufgaben im Fach Deutsch insgesamt als etwas weniger anspruchsvoll eingeschätzt als die landeseigenen Aufgaben. Zudem beurteilten sie die Erwartungshorizonte der Poolaufgaben als etwas weniger nützlich als diejenigen der landeseigenen Aufgaben.
- Im Fach Englisch zeigten sich vor allem bei den Poolaufgaben zum Kompetenzbereich „Hörverstehen“ Auffälligkeiten. Hier haben die Prüflinge insgesamt etwas bessere Ergebnisse erzielt als bei den übrigen landeseigenen Aufgaben.
- Gleichzeitig haben die Lehrkräfte die Poolaufgaben zum „Hörverstehen“ im Fach Englisch insgesamt im Vergleich zu den landeseigenen Aufgaben als anspruchsvoller eingeschätzt.
- Auch im Fach Französisch wurden für die Poolaufgaben zum Kompetenzbereich „Hörverstehen“ Auffälligkeiten festgestellt. Hier haben die Prüflinge insgesamt weniger gute Ergebnisse erzielt als bei den übrigen landeseigenen Aufgaben. Gleichzeitig wurden die Poolaufgaben zum Kompetenzbereich „Hörverstehen“ im Vergleich zu den Kompetenzbereichen „Schreiben“ und „Sprachmittlung“ von den Lehrkräften als etwas anspruchsvoller eingeschätzt.
- Die als Indikatoren zur kriterialen Validität der Poolaufgaben im Fach Französisch berechneten Kennwerte deuten zudem darauf hin, dass die in den Poolaufgaben erzielten Leistungen in den Kompetenzbereichen „Schreiben“ und „Sprachmittlung“ recht gut die bereits in der Qualifikationsphase erbrachten Ergebnisse abbilden. Deutlich geringer fällt hingegen der Zusammenhang zwischen den Prüfungsleistungen und den zugehörigen Vorleistungen im Kompetenzbereich „Hörverstehen“ aus.
- Im Fach Mathematik zeigte sich, dass die Prüflinge im Sachgebiet „Stochastik“ bei den Poolaufgaben insgesamt weniger gute Leistungen erzielten als bei den landeseigenen Aufgaben.
- Außerdem weisen im Fach Mathematik die als Indikatoren zur kriterialen Validität der Poolaufgaben ermittelten Kennwerte darauf hin, dass die in den Poolaufgaben erzielten Leistungen in den Sachgebieten „Analysis“ und „Analytische Geometrie/Lineare Algebra“ recht gut die bereits in der Qualifikationsphase erbrachten Ergebnisse abbilden. Weniger hoch fällt hingegen der Zusammenhang zwischen den Prüfungsleistungen und den zugehörigen Vorleistungen im Sachgebiet „Stochastik“ aus.

2 Methodisches Vorgehen

Mit Beschluss der „Konzeption zur Implementation der Bildungsstandards für die Allgemeine Hochschulreife“ hat die KMK das IQB am 10.10.2013 beauftragt, die Entwicklung und Nutzung der gemeinsamen Abituraufgabenpools der Länder auch wissenschaftlich zu begleiten. Auf dieser Grundlage wurde der erstmalige Einsatz von Aufgaben aus den Abituraufgabenpools im Prüfungsjahr 2017 formativ evaluiert.

2.1 Kernpunkte des Evaluationskonzepts

Die formative Evaluation basierte auf einem mit den Ländern abgestimmten Evaluationskonzept. Den Kern dieses Konzepts bildeten die folgenden sechs fächerübergreifenden Evaluationsbereiche:

(1) Auswahl der Aufgaben

- Wie häufig werden die Aufgaben aus den Pools im Vergleich zu den landeseigenen Aufgaben gewählt (sofern Wahlmöglichkeiten bestehen)?
- Gibt es einen Zusammenhang zwischen der Auswahl der Aufgaben und bestimmten Hintergrundmerkmalen der Prüflinge (z. B. Geschlecht, Vorleistungen in der Qualifikationsphase)?

(2) Schwierigkeit der Aufgaben

- Wie erfolgreich werden die Aufgaben aus den Pools bearbeitet?
- Unterscheidet sich die Schwierigkeit der Aufgaben aus den Pools von der Schwierigkeit landeseigener Aufgaben?
- Unterscheiden sich die Aufgaben aus den Pools und die landeseigenen Aufgaben in ihrer Schwierigkeit auch dann, wenn die Vorleistungen der Prüflinge berücksichtigt werden?

(3) Trennschärfe der Aufgaben

- Unterscheidet sich die Trennschärfe der Aufgaben aus den Pools von der Trennschärfe landeseigener Aufgaben?

(4) Validität der Aufgaben

- Gibt es einen Zusammenhang zwischen den Ergebnissen, die bei den Aufgaben aus den Pools erzielt wurden, und den Vorleistungen der Prüflinge?

(5) Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben

- Wie schätzen Lehrkräfte den Anspruch der Aufgaben ein?
- Wie verhält sich die Einschätzung zum Anspruch der Aufgaben aus den Pools zur Einschätzung zum Anspruch der landeseigenen Aufgaben?

(6) Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

- Wie schätzen die Lehrkräfte den Nutzen der Erwartungshorizonte (EWH) ein?
- Wie verhält sich die Einschätzung zum Nutzen der EWH der Aufgaben aus den Pools zur Einschätzung zum Nutzen der EWH der landeseigenen Aufgaben?

2.2 Stichprobenziehung und Datenerhebung

Die gemeinsamen Abituraufgabenpools der Länder umfassen Aufgaben zu den Fächern Deutsch und Mathematik sowie Englisch und Französisch als fortgeführte Fremdsprachen. Diejenigen Aufgaben der Pools für das Prüfungsjahr 2017, die von den Ländern entnommen wurden, können auf den Internetseiten des IQB unter www.iqb.hu-berlin.de → *Abituraufgaben* eingesehen werden. [1]

Zur Evaluation der Aufgaben der Pools wurden in jedem Land an zufällig ausgewählten allgemeinbildenden Schulen mit gymnasialer Oberstufe Daten erhoben. Zur Auswahl der Schulen wurden die Länder gebeten, dem IQB bis Ende Februar 2017 anonymisierte Listen aller in Frage kommenden Schulen zur Verfügung zu stellen. Auf dieser Grundlage erfolgte die Ziehung der Schulstichproben, die pro Land 20 Schulen umfassten. Dabei wurde in jedem Land darauf geachtet, dass das Verhältnis zwischen Gymnasien und anderen allgemeinbildenden Schulen mit einer gymnasialen Oberstufe (z. B. Gesamtschulen) in der Stichprobe dem entsprechenden Verhältnis in der Gesamtheit aller Schulen entspricht.

An jeder Schule der Stichproben wurde für jedes der vier Fächer und für jedes angebotene Anforderungsniveau ein Kurs in die Datenerhebung einbezogen, sofern die betreffende Abiturprüfung Aufgaben des Pools enthielt. Damit waren maximal 8 Kurse pro Schule an der Datenerhebung beteiligt. Die Auswahl dieser Kurse erfolgte durch die jeweilige Schulleitung. Um Lehrkräfte großer Kurse zu entlasten, wurde die Anzahl der Abiturientinnen und Abiturienten pro Kurs, deren Prüfungsergebnisse eingegeben werden sollten, auf 20 beschränkt.[2]

In den meisten Ländern erfolgte die Datenerhebung durch ein vom IQB entwickeltes webbasiertes Eingabeinstrument. Neben Informationen zur Auswahl der Aufgaben und zur Bewertung der Prüfungsarbeiten wurden dabei auch Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben und zum Nutzen des EWH bei der Bewertung der Prüfungsarbeiten erfasst. Die Lehrkräfte nahmen diese Einschätzungen anhand einer fünfstufigen Skala vor und erhielten darüber hinaus die Möglichkeit, sich in einem Kommentarfeld frei zu äußern. Dabei war ihnen nicht bekannt, welche der Aufgaben aus den gemeinsamen Abituraufgabenpools der Länder stammen. Um Zusammenhänge zwischen der Bearbeitung der Aufgaben und Hintergrundmerkmalen der Abiturientinnen und Abiturienten zu untersuchen, wurden zusätzlich Daten zu ihrem Geschlecht, ihrem Sprachhintergrund und ihren Vorleistungen (Klausur- und Halbjahresnoten) in der Qualifikationsphase erhoben.

2.3 Überblick zur Stichprobe

Die im vorliegenden Bericht dargestellten Evaluationsergebnisse basieren auf den Datensätzen aus 14 Ländern, die nach Abschluss der Datenerhebung am 31.12.2017 vorlagen. Angaben zur Anzahl der in die Evaluation einbezogenen Kurse und Prüflinge können Tabelle 1 entnommen werden.

[1] Sind für eine Aufgabe Nutzungsrechte für zugrunde liegende Materialien erforderlich, so wird diese nur veröffentlicht, wenn die Nutzungsrechte erworben werden können.

[2] Die Auswahl der Schülerinnen und Schüler erfolgte durch die zuständigen Lehrkräfte. Die Auswahl sollte so erfolgen, dass ein möglichst breites Leistungsspektrum abgebildet wird. Vermieden werden sollte eine selektive Berücksichtigung bzw. Nichtberücksichtigung bestimmter Gruppen (z. B. besonders leistungsschwache oder leistungsstarke Prüflinge, Schülerinnen und Schüler mit nichtdeutscher Herkunftssprache).

Tabelle 1: Umfang der Kurs- bzw. Schülerstichproben

	Deutsch	Englisch	Französisch	Mathematik
Anzahl der Kurse	272	413	297	477
Anzahl der Prüflinge	5167	6602	1852	8094
… auf erhöhtem Niveau	4312	5385	1659	5352
… auf grundlegendem Niveau	855	1217	193	2742

2.4 Datenauswertung

Die zur Bewährung der einzelnen Aufgaben erhobenen Daten wurden mittels metaanalytischer Methoden zu länderübergreifenden Gesamtergebnissen zusammengefasst. Dabei wird jedes Einzelergebnis (d. h. jeder Kennwert, der für eine Aufgabe aus dem Pool in einem einzelnen Land berechnet wurde) als eine Einheit in die Auswertung einbezogen. Zudem wurden verschiedene Differenzierungen vorgenommen (z. B. nach Kompetenzbereichen, Sachgebieten oder Aufgabenarten).[3] Als Ergebnis der Metaanalysen wird jeweils ein über alle Länder, Anforderungsniveaus und Poolaufgaben zusammengefasster Effekt berechnet. Dabei wurde in allen Ergebnisdarstellungen ein Vergleich zwischen den aus den Pools eingesetzten Aufgaben einerseits und den landeseigenen Aufgaben, die nicht Teil der Pools waren, andererseits vorgenommen. Tabelle 2 ist zu entnehmen, in welcher Hinsicht die Vergleiche für jedes Fach durchgeführt wurden.

Tabelle 2: Überblick zu den in den Ergebnisdiagrammen vorgenommenen Vergleichen

Fach	Vergleich
Deutsch	Die Ergebnisse zu den Poolaufgaben werden mit dem Mittelwert der Ergebnisse zu allen landeseigenen Aufgaben verglichen. Dabei sind Besonderheiten in der Zusammensetzung der Aufgaben aus dem Pool zu beachten: Unter den eingesetzten Poolaufgaben sind (im Vergleich zu den landeseigenen Aufgaben) die Aufgaben zum materialgestützten Schreiben überrepräsentiert und insbesondere textbezogene Aufgaben zu literarischen Texten unterrepräsentiert.
Englisch und Französisch	Die Ergebnisse zu den Poolaufgaben (z. B. zum Kompetenzbereich „Hörverstehen“) werden mit dem Mittelwert der Ergebnisse zu allen landeseigenen Aufgaben (also z. B. auch zu Aufgaben zu den Kompetenzbereichen „Schreiben“ und „Sprachmittlung“) verglichen.
Mathematik	Die Ergebnisse der aus dem Pool eingesetzten Teilaufgaben werden zusammengefasst und mit dem Mittelwert der Ergebnisse zu allen landeseigenen Teilaufgaben verglichen. Zudem erfolgt eine Differenzierung nach Sachgebieten.

[3] Dazu wurden sogenannte Random-Effects-Metaanalysen mit einem bayesianischen Schätzverfahren und Hartung-Knapp Korrektur berechnet.

2.5 Erläuterungen zur Interpretation der Ergebnistabellen

Differenzwerte

Die Ergebnisse der Metaanalysen für die Bereiche „Schwierigkeit der Aufgaben“, „Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben“ und „Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben“ werden jeweils als Differenzen zwischen den für die Poolaufgaben ermittelten Werten und den Ergebnissen für die landeseigenen Aufgaben dargestellt. Eine positive Differenz bedeutet, dass die Prüfungsergebnisse bzw. die Lehrkräfteeinschätzungen bei den Poolaufgaben besser ausfielen als bei den landeseigenen Aufgaben. Eine negative Differenz weist hingegen auf weniger gute Ergebnisse bzw. Einschätzungen bei den Poolaufgaben hin. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert. Für jeden dargestellten Differenzwert wird zusätzlich ein Vertrauensbereich genannt. Dieser gibt an, in welchem Bereich sich der ermittelte Wert mit einer Wahrscheinlichkeit von 95 % befindet, wenn die Datenerhebung mit einer anderen Stichprobe wiederholt werden würde.

Die Differenzwerte werden entweder in Form von Notenpunkten (NP) oder aber standardisiert dargestellt. Als standardisiertes Effektmaß wurde Hedges’ g bestimmt. Dieses Effektmaß ist wie folgt zu interpretieren: Effekte von $g < 0{,}20$ gelten als sehr gering und können in der Regel vernachlässigt werden. Effekte ab $g = 0{,}20$ werden zumeist als „klein“ eingestuft, Effekte ab $g = 0{,}50$ gelten als „mittel“ und Effekte ab $g = 0{,}80$ können als „hoch“ bewertet werden (z. B. Borenstein et al., 2009).

Trennschärfekoeffizienten

Die Trennschärfe einer Aufgabe wurde über die Korrelation der bei dieser Aufgabe erzielten Ergebnisse mit den Gesamtergebnissen der Abiturprüfung berechnet. Allgemein können Trennschärfekoeffizienten zwischen $r = 0{,}30$ und $r = 0{,}50$ als „mittel“ eingestuft werden, Kennwerte über $r = 0{,}50$ gelten als hoch und Kennwerte unter $r = 0{,}20$ werden als gering bewertet (z. B. Bortz & Döring, 2006). Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert. Zusätzlich ist zu jedem Trennschärfekoeffizienten der Vertrauensbereich angegeben.

Validitätskoeffizienten

Als Indikator für die kriteriale Validität einer Aufgabe wurden die Korrelationen der jeweils bei dieser Aufgabe erzielten Ergebnisse mit den in der Qualifikationsphase erreichten Klausur- bzw. Halbjahresnoten im betreffenden Fach bestimmt. Die Höhe der angegebenen Werte kann wie folgt interpretiert werden: Validitätskoeffizienten unter $r = 0{,}40$ werden in der Forschungsliteratur häufig als „klein“ bewertet. Koeffizienten von $r = 0{,}40$ bis $r = 0{,}60$ sind als „mittel“ einzustufen, Validitätskennwerte ab $r = 0{,}60$ gelten als „hoch“ (vgl. Fisseni, 1997). Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert. Zusätzlich ist zu jedem Validitätskoeffizienten der Vertrauensbereich angegeben.

3 Ergebnisse zur Bewährung der Aufgaben aus den Pools

Im Folgenden werden für jedes der vier Fächer die Ergebnisse zur Bewährung der Aufgaben in Bezug auf die sechs im Evaluationskonzept spezifizierten Bereiche (vgl. Abschnitt 2.1) zusammengefasst.

3.1 Deutsch

3.1.1 Auswahl der Aufgaben

Bei der schriftlichen Abiturprüfung im Fach Deutsch können die Prüflinge in allen Ländern aus mindestens zwei Alternativen eine Abiturprüfungsaufgabe auswählen.[4] Die Ergebnisse zur Auswahlhäufigkeit der Aufgaben lassen sich wie folgt zusammenfassen:

- Die meisten der aus dem Abituraufgabenpool stammenden Aufgaben wurden von relativ vielen Prüflingen und vielfach sogar etwas häufiger als die landeseigenen Aufgaben gewählt.
- Die Aufgabe „Damals“ („Interpretation literarischer Texte“) ist in einem Land nur von vier Prozent der Schülerinnen und Schüler bearbeitet worden. Allerdings wurde die Aufgabe in allen anderen Ländern, die sie eingesetzt haben, deutlich häufiger ausgewählt. Ebenfalls relativ selten wurde die Aufgabe „Reise“ („materialgestütztes Verfassen informierender Texte“) bearbeitet.
- Die Aufgaben der Art „materialgestütztes Verfassen argumentierender Texte“ wurden länderübergreifend insgesamt relativ häufig von den Schülerinnen und Schülern ausgewählt.

In einer Metaanalyse wurde aufgabenübergreifend der Zusammenhang zwischen der Auswahl der Aufgaben und bestimmten Hintergrundmerkmalen der Prüflinge betrachtet. Unter anderem wurde untersucht, ob eine bestimmte Aufgabenart eher von Schülerinnen oder eher von Schülern gewählt wurde und inwieweit eher Prüflinge, die in der Qualifikationsphase gute Vorleistungen gezeigt haben, oder jene, die weniger gute Vorleistungen erbracht haben, die jeweilige Aufgabenart gewählt haben. Die Ergebnisse lassen sich wie folgt zusammenfassen:

- Die Aufgabenart „Interpretation literarischer Texte“ wurde sehr viel häufiger von Schülerinnen gewählt, während sich Schüler häufiger für die Aufgabenart „Analyse pragmatischer Texte“ entschieden.
- Die Ergebnisse zur Auswahlhäufigkeit (vgl. Tabelle 3) der Poolaufgaben legen ferner insgesamt den Schluss nahe, dass textbezogene Aufgaben zu literarischen Texten eher von Schülerinnen und Schülern mit guten Vorleistungen bzw. Halbjahresnoten gewählt werden, während sich Prüflinge mit weniger guten Vorleistungen offenbar eher für Aufgaben zu pragmatischen Texten (hier: „Analyse pragmatischer Texte“, „Erörterung pragmatischer Texte“) oder für Aufgaben zum „materialgestützten Verfassen argumentierender/informierender Texte“ entscheiden. Da jedoch die textbezogenen Aufgaben zu literarischen Texten bei den aus dem Pool eingesetzten Aufgaben deutlich unterrepräsentiert waren, wurde für die Poolaufgaben insgesamt festgestellt, dass diese eher von Prüflingen gewählt wurden, die in der Qualifikationsphase weniger gute Vorleistungen bzw. Halbjahresnoten zeigten (Differenz: -0.4 NP, statistisch signifikant).

[4] In einigen Ländern nehmen die zuständigen Lehrkräfte zusätzlich eine Vorauswahl vor.

Tabelle 3: Aufgabenübergreifende Ergebnisse zur Auswahl der Aufgaben und den Halbjahresleistungen in der Qualifikationsphase im Fach Deutsch

Aufgabenart[5]	Anzahl Länder	Differenz (in NP)	Vertrauensbereich (95 %)
AP	2	-0,83	-1,40 / -0,26
EP	1	-0,11	-1,00 / 0,78
IL	7	0,14	-0,37 / 0,65
MA	8	-0,64	-1,02 / -0,25
MI	1	-1,07	-1,85 / -0,28
insgesamt	**19**	**-0,40***	**-0,71 / -0,10**

3.1.2 Schwierigkeit der Aufgaben

Als Indikator für die empirische Aufgabenschwierigkeit der aus dem Pool im Fach Deutsch eingesetzten Aufgaben wurden jeweils die arithmetischen Mittelwerte der von den Prüflingen erzielten Notenpunkte berechnet. Zudem wurde untersucht, ob sich die aus dem Pool eingesetzten Aufgaben hinsichtlich der Aufgabenschwierigkeit von den landeseigenen Aufgaben unterscheiden. Insgesamt lassen sich die dazu ermittelten Ergebnisse wie folgt zusammenfassen:

- Die Ergebnisse der Metaanalyse (vgl. Tabelle 4) zeigen, dass die im Fach Deutsch erreichten Notenpunkte bei den Poolaufgaben insgesamt etwas geringer sind als bei den landeseigenen Aufgaben (Differenz: -0,39 NP, statistisch signifikant). Wie bereits unter 3.1.1 skizziert, ist dieses Gesamtergebnis offenbar dadurch bedingt, dass die Poolaufgaben, mit Ausnahme der Aufgaben zur „Interpretation literarischer Texte“, häufiger von Schülerinnen und Schülern mit weniger guten Vorleistungen gewählt wurden.
- Übereinstimmend mit dieser Auslegung ist der für die Aufgaben der Art „Interpretation literarischer Texte“ festgestellte Mittelwert der Notenpunkte insgesamt sogar etwas höher als bei den jeweils zum Vergleich herangezogenen landeseigenen Aufgaben (Differenz: +0,14 NP, nicht statistisch signifikant).
- Ein im Vergleich zu den landeseigenen Aufgaben signifikant geringerer Mittelwert der Notenpunkte wurde hingegen für die Aufgaben der Art „Analyse pragmatischer Texte“ (zwei Aufgaben auf grundlegendem Niveau, Differenz: -0,79 NP) und für die Aufgabe der Art „materialgestütztes Verfassen informierender Texte“ (eine Aufgabe auf erhöhtem Niveau, Differenz: -1,97 NP) festgestellt. Allerdings ist das für die letztgenannte Aufgabenart ermittelte Ergebnis mit einer hohen statistischen Unsicherheit behaftet, da die betreffende Aufgabe nur von wenigen Prüflingen gewählt wurde und die Datenbasis daher klein ist.
- Mit einer Ausnahme (Aufgabe „Straße“) ist für alle Aufgaben zum materialgestützten Schreiben im Durchschnitt ein etwas geringerer Notenpunktwert als bei den landeseigenen Vergleichsaufgaben ermittelt worden. In einigen Fällen sind diese Unterschiede auch statistisch signifikant (z. B. für die Aufgabe „Medien“). Dies kann jedoch zumeist darauf zurückgeführt werden, dass die Aufgaben zum materialgestützten Schreiben häufiger von Schülerinnen und Schülern mit weniger guten Vorleistungen im Fach Deutsch ausgewählt wurden.

[5] verwendete Abkürzungen: AP- Analyse pragmatischer Texte, EP - Erörterung pragmatischer Texte, IL - Interpretation literarischer Texte, MA - materialgestütztes Verfassen argumentierender Texte, MI - materialgestütztes Verfassen informierender Texte

Tabelle 4: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Deutsch

Aufgabenart[6]	Anzahl Länder	Differenz (in NP)	Vertrauensbereich (95 %)
AP	2	-0,79	-0,98 / -0,60
EP	1	-0,29	-1,45 / 0,86
IL	7	0,14	-0,52 / 0,81
MA	8	-0,48	-1,13 / 0,17
MI	1	-1,97	-3,24 / -0,71
insgesamt	**19**	**-0,39***	**-0,78 / -0,01**

3.1.3 Trennschärfe der Aufgaben

Die Trennschärfe einer Aufgabe ist als Korrelation dieser Aufgabe mit dem Gesamtergebnis eines Tests bzw. einer Prüfung definiert. Im Fach Deutsch lässt sich dieser Kennwert nicht sinnvoll bestimmen, da das bei einer Aufgabe erzielte Ergebnis jeweils identisch mit dem Prüfungsergebnis ist.

3.1.4 Validität der Aufgaben

Als Indikator für die kriteriale Validität wurden Korrelationen zwischen den bei den jeweiligen Aufgaben erzielten Ergebnissen einerseits und den in der Qualifikationsphase erreichten Klausur- bzw. Halbjahresnoten andererseits bestimmt. Die dazu ermittelten Ergebnisse sind in den Tabellen 5 und 6 dargestellt und lassen sich wie folgt zusammenfassen:

- In Bezug auf das Validitätskriterium „Halbjahresnoten im Fach Deutsch“ wurde mit Metaanalysen für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von $r = 0{,}69$ ermittelt. Dieser Koeffizient unterscheidet sich zwar statistisch signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde ($r = 0{,}74$), beide Koeffizienten sind jedoch als hoch einzustufen. Auch die für die einzelnen Aufgabenarten berechneten Validitätskoeffizienten sind nahezu konsistent als hoch zu bewerten (zwischen $r = 0{,}66$ und $r = 0{,}80$). Einzig der Kennwert für die Aufgabenart „materialgestütztes Verfassen informierender Texte“, der allerdings nur für eine relativ kleine Stichprobe bestimmt werden konnte, ist mit $r = 0{,}51$ etwas geringer.
- In Bezug auf das Validitätskriterium „Klausurnoten im Fach Deutsch“ wurde metaanalytisch für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von $r = 0{,}72$ ermittelt. Dieser Koeffizient unterscheidet sich nicht statistisch signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde ($r = 0{,}77$). Beide Koeffizienten sind als hoch einzustufen. Auch die für die einzelnen Aufgabenarten berechneten Validitätskoeffizienten sind jeweils als hoch zu bewerten (zwischen $r = 0{,}60$ bis $r = 0{,}83$).

[6] verwendete Abkürzungen: AP - Analyse pragmatischer Texte, EP - Erörterung pragmatischer Texte, IL - Interpretation literarischer Texte, MA - materialgestütztes Verfassen argumentierender Texte, MI - materialgestütztes Verfassen informierender Texte

Tabelle 5: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase im Fach Deutsch

Aufgabenart[7]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
AP	2	0,73	0,67 / 0,78
EP	1	0,80	0,70 / 0,87
IL	7	0,70	0,58 / 0,79
MA	8	0,66	0,61 / 0,70
MI	1	0,51	0,11 / 0,77
insgesamt	**19**	**0,69***	**0,65 / 0,72**
landeseigene Aufgaben			
insgesamt	**37**	**0,74***	**0,72 / 0,76**

Tabelle 6: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase im Fach Deutsch

Aufgabenart[6]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
AP	2	0,73	0,68 / 0,68
EP	1	0,83	0,74 / 0,89
IL	7	0,76	0,96 / 0,81
MA	8	0,69	0,62 / 0,75
MI	1	0,83	0,74 / 0,89
insgesamt	**19**	**0,72***	**0,68 / 0,76**
landeseigene Aufgaben			
insgesamt	**37**	**0,77***	**0,74 / 0,79**

3.1.5 Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben

Metaanalytisch lassen sich die Ergebnisse für die Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben im Fach Deutsch anhand von standardisierten Effektmaßen (hier: Hedges' *g*) wie folgt zusammenfassen (vgl. Tabelle 7):

- Insgesamt haben die Lehrkräfte die Aufgaben aus dem Pool als etwas weniger anspruchsvoll eingeschätzt als die landeseigenen Aufgaben. Der Unterschied ist allerdings statistisch nicht signifikant und zudem sehr gering (g = -0,12).
- Auch bei einer nach Aufgabenarten differenzierten Betrachtung finden sich keine statistisch signifikanten Unterschiede im Anspruch zwischen den Poolaufgaben und den landeseigenen Aufgaben.

[7] verwendete Abkürzungen: AP - Analyse pragmatischer Texte, EP - Erörterung pragmatischer Texte, IL - Interpretation literarischer Texte, MA - materialgestütztes Verfassen argumentierender Texte, MI - materialgestütztes Verfassen informierender Texte

- Bei einer Analyse auf Aufgabenebene zeigt sich allerdings, dass die Aufgabe „Medien“ („materialgestütztes Verfassen argumentierender Texte“) im Vergleich zu den jeweiligen landeseigenen Aufgaben signifikant weniger anspruchsvoll eingeschätzt wurde. Demgegenüber wurde der Anspruch der Aufgabe „Hallo“ („Analyse pragmatischer Texte“) signifikant höher bewertet als der Anspruch landeseigener Aufgaben.

Tabelle 7: Aufgabenübergreifende Ergebnisse zum eingeschätzten Anspruch der Aufgaben im Fach Deutsch

Aufgabenart[8]	Anzahl Länder	Stand. Differenz (*g*)	Vertrauensbereich (95 %)
AP	2	0,24	0,06 / 0,41
EP	1	-0,12	-0,76 / 0,52
IL	7	-0,22	-0,58 / 0,15
MA	8	-0,11	-0,38 / 0,15
MI	1	-0,31	-0,93 / 0,31
insgesamt	**19**	**-0,12**	**-0,28 / 0,04**

3.1.6 Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

Die metaanalytische Zusammenfassung der Ergebnisse zur Einschätzung der Lehrkräfte zum Nutzung der EWH der Aufgaben ergibt folgendes Befundmuster (vgl. Tabelle 8):

- Insgesamt haben die Lehrkräfte den Nutzen der EWH bei den Aufgaben aus dem Pool signifikant geringer eingeschätzt als bei den landeseigenen Aufgaben, wobei der Unterschied jedoch klein ist (g = -0,28).
- Eine nach Aufgabenarten differenzierte Betrachtung der Ergebnisse zeigt zudem, dass diese Diskrepanz vor allem auf die weniger positiven Einschätzungen der Lehrkräfte zur Nützlichkeit der EWH bei den Aufgaben zum „materialgestützten Verfassen argumentierender Texte“ (g = -0,50, mittlerer Effekt, statistisch signifikant) und bei den Aufgaben zur „Analyse pragmatischer Texte“ (g = -0,50, mittlerer Effekt, allerdings nur für die Aufgabe „Hallo“ statistisch signifikant) zurückzuführen ist.

Tabelle 8: Aufgabenübergreifende Ergebnisse zum eingeschätzten Nutzen der EWH der Aufgaben im Fach Deutsch

Aufgabenart[7]	Anzahl Länder	Stand. Differenz (*g*)	Vertrauensbereich (95 %)
AP	2	-0,50	-2,05 / 1,05
EP	1	-0,16	-0,79 / 0,48
IL	7	0,11	-0,08 / 0,30
MA	8	-0,50	-0,83 / -0,17
MI	1	-0,38	-1,00 / 0,25
insgesamt	**19**	**-0,28***	**-0,47 / -0,08**

[8] verwendete Abkürzungen: AP - Analyse pragmatischer Texte, EP - Erörterung pragmatischer Texte, IL - Interpretation literarischer Texte, MA - materialgestütztes Verfassen argumentierender Texte, MI - materialgestütztes Verfassen informierender Texte

3.2 Englisch

3.2.1 Auswahl der Aufgaben

In den schriftlichen Abiturprüfungen der Länder im Fach Englisch haben die Prüflinge in der Regel weniger Wahlmöglichkeiten als im Fach Deutsch. In einigen Ländern können sie in den Kompetenzbereichen „Schreiben“ und „Sprachmittlung“ zwischen (zumeist) zwei Aufgaben bzw. Aufgabenblöcken wählen. Im Kompetenzbereich „Hörverstehen“ ist es grundsätzlich nicht möglich, Wahlmöglichkeiten anzubieten, da die Schülerinnen und Schüler die Aufgaben jeweils gleichzeitig bearbeiten.

Insgesamt zeigen die Ergebnisse, dass die Aufgaben aus dem Pool im Vergleich zu den landeseigenen Aufgaben ähnlich häufig bzw. tendenziell sogar etwas häufiger gewählt wurden. Allerdings variieren die Werte für die Auswahlhäufigkeit der Aufgaben deutlich zwischen den Ländern.

Aufgrund der eingeschränkten Wahlmöglichkeiten im Fach Englisch ist es für viele Länder nicht möglich, Vergleiche zwischen Poolaufgaben und landeseigenen Aufgaben desselben Kompetenzbereichs durchzuführen. Aus diesem Grund wurden im Fach Englisch die Ergebnisse zu den aus dem Pool eingesetzten Aufgaben (z. B. zum Kompetenzbereich „Hörverstehen“) jeweils mit dem Mittelwert der Ergebnisse zu allen landeseigenen Aufgaben (also z. B. auch zu Aufgaben zu den Kompetenzbereichen „Sprachmittlung“ und „Schreiben“) verglichen.

3.2.2 Schwierigkeit der Aufgaben

Wie für das Fach Deutsch wurde auch für das Fach Englisch im Evaluationsbereich „Schwierigkeit der Aufgaben“ ermittelt, wie erfolgreich die Prüflinge die Aufgaben aus den Pool bearbeitet haben und welche Unterschiede zu den anderen Aufgaben der jeweiligen Abiturprüfung in den Ländern bestehen.

Die im Fach Englisch zur Aufgabenschwierigkeit berechneten Kennwerte wurden mittels metaanalytischer Methoden zu standardisierten Effektmaßen aggregiert. Die resultierenden Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Tabelle 9):

- Insgesamt unterscheiden sich die bei den Aufgaben aus dem Pool von den Schülerinnen und Schülern im Mittel erzielten Ergebnisse nicht statistisch signifikant von den Ergebnissen bei den landeseigenen Aufgaben (g = 0,08, sehr geringer Effekt).
- Bei einer nach Kompetenzbereichen differenzierten Betrachtung der Prüfungsergebnisse zeigt sich jedoch, dass die Schülerinnen und Schüler bei den Poolaufgaben zum „Hörverstehen“ insgesamt signifikant höhere Notenpunkte bzw. Lösungsquoten erreicht haben als bei den landeseigenen Aufgaben (g = 0,27, kleiner Effekt).
- Gleichzeitig finden sich jedoch auch Länder, deren Ergebnisse beim “Hörverstehen“ von diesem übergreifenden Befundmuster abweichen. Dass die aus dem Abituraufgabenpool stammenden Aufgaben zum Kompetenzbereich „Hörverstehen“ in vielen Ländern im Vergleich zu den anderen Prüfungsaufgaben für die Schülerinnen und Schüler etwas leichter, in einigen Ländern hingegen etwas schwieriger sind, hängt möglicherweise damit zusammen, dass Prüfungsaufgaben zu diesem Kompetenzbereich in den Ländern eine unterschiedlich lange Tradition haben und in einigen Ländern sogar erstmalig in der schriftlichen Abiturprüfung eingesetzt wurden.

- In den übrigen beiden Kompetenzbereichen („Schreiben“ und „Sprachmittlung“) wurden keine statistisch signifikanten Unterschiede zwischen den in Poolaufgaben und den in landeseigenen Aufgaben erzielten Ergebnissen festgestellt (g = -0,13 bzw. g = 0,05, jeweils sehr geringe Effekte).

Tabelle 9: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Englisch

KB[9]	Anzahl Länder	Stand. Differenz (g)	Vertrauensbereich (95 %)
HV	9	0,27	0,01 / 0,53
SM	15	0,05	-0,09 / 0,20
S	6	-0,13	-0,59 / 0,33
insgesamt	**30**	**0,08**	**-0,05 / 0,21**

3.2.3 Trennschärfe der Aufgaben

Die Trennschärfe einer Aufgabe gibt an, wie gut sie das Gesamtergebnis der Prüfung repräsentiert. Der Trennschärfekoeffizient einer Aufgabe wird als Korrelation zwischen den bei der Aufgabe erzielten Ergebnissen mit den Gesamtergebnissen berechnet (Bortz & Döring, 2006). Die metaanalytische Zusammenfassung der Ergebnisse zur Trennschärfe der Aufgaben im Fach Englisch ergibt folgendes Muster (vgl. Tabelle 10):

- Der für alle Aufgaben aus dem Pool im Fach Englisch aggregierte Trennschärfekoeffizient ist als hoch einzustufen (r = 0,86) und unterscheidet sich nicht statistisch signifikant von dem für die landeseigenen Aufgaben berechneten Koeffizienten (r = 0,88).
- Für die Poolaufgaben zum „Hörverstehen“ wurde ein aggregierter Trennschärfekoeffizient von r = 0,74 ermittelt. Dieser Koeffizient ist signifikant kleiner als der Kennwert für die Poolaufgaben zum „Schreiben“ (r = 0,92), unterscheidet sich jedoch nicht signifikant vom Koeffizienten für die Poolaufgaben zur „Sprachmittlung“ (r = 0,87).

Tabelle 10: Aufgabenübergreifende Ergebnisse zur Trennschärfe im Fach Englisch

KB[8]	Anzahl Länder	Trennschärfekoeffizient r	Vertrauensbereich (95 %)
Poolaufgaben			
HV	11	0,74	0,67 / 0,80
SM	16	0,87	0,85 / 0,90
S	8	0,92	0,82 / 0,97
insgesamt	**35**	**0,86***	**0,82 / 0,89**
landeseigene Aufgaben			
insgesamt	**51**	**0,88***	**0,84 / 0,91**

[9] verwendete Abkürzungen: KB - Kompetenzbereich, HV - Hörverstehen. S - Schreiben, SM - Sprachmittlung

3.2.4 Validität der Aufgaben

Die als Indikatoren zur kriterialen Validität der Aufgaben im Fach Englisch ermittelten Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Tabellen 11 und 12):

- In Bezug auf das Validitätskriterium „Halbjahresnoten“ wurde für die Aufgaben aus dem Pool ein aggregierter Koeffizient von $r = 0{,}70$ ermittelt. Dieser unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde ($r = 0{,}71$). Beide Koeffizienten sind als hoch einzustufen. Bei einer nach Kompetenzbereichen differenzierten Analyse findet sich der höchste Validitätskoeffizient für die Aufgaben zum „Schreiben“ ($r = 0{,}75$). Auch für die Sprachmittlung ist der Koeffizient mit $r = 0{,}72$ hoch ausgeprägt. Dieser Wert ist signifikant höher als der für die Aufgaben zum „Hörverstehen“ ermittelte Koeffizient ($r = 0{,}63$).
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von $r = 0{,}72$ ermittelt. Dieser unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde ($r = 0{,}71$). Während die für die Poolaufgaben zum „Schreiben“ ($r = 0{,}79$) und zur „Sprachmittlung“ ($r = 0{,}76$) gefundenen Validitätskoeffizienten nahezu identisch sind, fällt der für die Poolaufgaben zum „Hörverstehen“ berechnete Kennwert ($r = 0{,}59$) wiederum signifikant geringer aus.

Tabelle 11: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase im Fach Englisch

KB[10]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
HV	11	0,63	0,59 / 0,66
SM	16	0,72	0,68 / 0,75
S	8	0,75	0,65 / 0,75
insgesamt	**35**	**0,70***	**0,67 / 0,73**
landeseigene Aufgaben			
insgesamt	**51**	**0,71***	**0,68 / 0,75**

Tabelle 12: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase im Fach Englisch

KB[10]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
HV	11	0,59	0,47 / 0,68
SM	16	0,76	0,72 / 0,80
S	8	0,79	0,67 / 0,86
insgesamt	**35**	**0,72***	**0,67 / 0,76**
landeseigene Aufgaben			
insgesamt	**51**	**0,71***	**0,66 / 0,76**

[10] verwendete Abkürzungen: KB - Kompetenzbereich, HV – Hörverstehen, S - Schreiben, SM - Sprachmittlung

3.2.5 Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben

Metaanalytisch lassen sich die Ergebnisse für die Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben im Fach Englisch anhand von standardisierten Effektmaßen wie folgt zusammenfassen (vgl. Tabelle 13):

- Insgesamt wurden die Aufgaben aus dem Pool als etwas anspruchsvoller bewertet als die landeseigenen Aufgaben. Der Unterschied ist jedoch statistisch nicht signifikant und zudem als sehr gering einzustufen (g = 0,18).
- Bei einer nach Kompetenzbereichen differenzierten Betrachtung fällt allerdings auf, dass die Lehrkräfte lediglich die Poolaufgaben zum „Hörverstehen“ als deutlich anspruchsvoller eingeschätzt haben als die landeseigenen Aufgaben (g = 0,56, mittlerer bis starker Effekt, statistisch signifikant). Bemerkenswert ist dieser Befund vor allem vor dem Hintergrund der Ergebnisse zur Aufgabenschwierigkeit, die (trotz leicht abweichender Befundmuster in einigen Ländern) insgesamt den Schluss nahe legen, dass die Schülerinnen und Schüler recht gut in der Lage waren, die aus dem Aufgabenpool stammenden Aufgaben zum „Hörverstehen“ zu bewältigen.

Tabelle 13: Aufgabenübergreifende Ergebnisse zum eingeschätzten Anspruch der Aufgaben im Fach Englisch

KB[11]	Anzahl Länder	Mittlere stand. Differenz	Vertrauensbereich (95 %)
HV	9	0,56	0,16 / 0,96
SM	14	-0,02	-0,23 / 0,20
S	6	0,15	-0,46 / 0,75
insgesamt	**29**	**0,18**	**-0,01 / 0,38**

3.2.6 Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

Im Hinblick auf den Nutzen der EWH finden sich sowohl in der Gesamtschau als auch bei einer nach Kompetenzbereichen separierten Betrachtung nur geringe, statistisch nicht signifikante Unterschiede in den Einschätzungen der Lehrkräfte zwischen den Aufgaben aus dem Pool und den landeseigenen Aufgaben.

[11] verwendete Abkürzungen: KB - Kompetenzbereich, HV - Hörverstehen. S - Schreiben, SM - Sprachmittlung

3.3 Französisch

Im Fach Französisch legen bundesweit betrachtet deutlich weniger Schülerinnen und Schüler eine schriftliche Abiturprüfung ab als in den Fächern Deutsch, Englisch und Mathematik. Die im Folgenden dargestellten Ergebnisse basieren dementsprechend auf erheblich kleineren Stichproben als die Befunde zu den anderen drei Fächern und sind daher statistisch weniger belastbar.

3.3.1 Auswahl der Aufgaben

In der Regel sind die Abiturprüfungen im Fach Französisch analog zum Fach Englisch strukturiert. Wie in Englisch werden daher nachfolgend auch die Ergebnisse zu den Poolaufgaben in Französisch jeweils mit dem Mittelwert der Ergebnisse zu allen landeseigenen Aufgaben verglichen.

Zur Auswahl der aus dem Abituraufgabenpool stammenden Aufgaben liegen insgesamt nur wenige Ergebnisse vor. In einigen Fällen wurden die Aufgaben aus dem Pool deutlich häufiger gewählt als die landeseigenen Aufgaben (z. B. „Maupassant“ und „Tête“ aus dem Kompetenzbereich „Schreiben" sowie „Berlin“ aus dem Kompetenzbereich Sprachmittlung), in anderen Fällen zeigen sich nur geringfügige Unterschiede oder höhere Auswahlhäufigkeiten für die landeseigenen Aufgaben.

3.3.2 Schwierigkeit der Aufgaben

Die im Fach Französisch zur Aufgabenschwierigkeit berechneten Kennwerte wurden mittels metaanalytischer Methoden zu standardisierten Effektmaßen aggregiert. Die resultierenden Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Tabelle 14):

- Insgesamt unterscheiden sich die Ergebnisse, die von den Schülerinnen und Schülern im Mittel bei den Aufgaben aus dem Pool erzielt wurden, nicht statistisch signifikant von den Ergebnissen in den landeseigenen Aufgaben (g = -0,12, sehr geringer Effekt).
- Bei einer nach Kompetenzbereichen differenzierten Betrachtung der Prüfungsergebnisse zeigt sich, dass die Schülerinnen und Schüler bei den Poolaufgaben zum „Hörverstehen“ weniger gute Ergebnisse erzielt haben als bei den landeseigenen Aufgaben. Dieses Ergebnismuster weicht von den Befunden im Fach Englisch ab, wo die Schülerinnen und Schüler bei den Poolaufgaben zum „Hörverstehen“ bessere Leistungen erreicht haben als bei den landeseigenen Aufgaben. Der für die Poolaufgaben zum „Hörverstehen“ im Fach Französisch festgestellte Unterschied ist als klein bis mittel einzustufen (g = -0,44, statistisch signifikant). Korrespondierend mit diesem Befund wurde von den Lehrkräften im Kommentarfeld mehrfach angemerkt, dass die Aufgaben zum „Hörverstehen“ zu schwierig gewesen seien. Zudem wurde angeregt, zukünftig mehr Übungsaufgaben zum „Hörverstehen“ bereitzustellen.
- Gleichzeitig finden sich auch Länder, deren Ergebnisse beim „Hörverstehen“ von diesem übergreifenden Befundmuster abweichen: Zumindest bei einer Aufgabe („Mentorat“) haben die Prüflinge in einem Land signifikant besser abgeschnitten als bei den landeseigenen Aufgaben.
- In den anderen beiden Kompetenzbereichen („Schreiben“ und „Sprachmittlung“) wurden keine statistisch signifikanten Unterschiede zwischen den von Schülerinnen und Schülern in den Poolaufgaben und in den landeseigenen Aufgaben erzielten Ergebnissen festgestellt (g = -0,07 bzw. g = 0,07, jeweils sehr geringe Effekte).

Tabelle 14: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Französisch

KB[12]	Anzahl Länder	Mittlere stand. Differenz	Vertrauensbereich (95 %)
HV	10	-0,44	-0,87 / 0,00
SM	14	0,07	-0,05 / 0,19
S	7	-0,07	-0,26 / 0,12
insgesamt	**31**	**-0,12**	**-0,28 / 0,02**

3.3.3 Trennschärfe der Aufgaben

Die metaanalytische Zusammenfassung der Ergebnisse zur Trennschärfe der Aufgaben im Fach Französisch ergibt folgendes Muster (vgl. Tabelle 15):

- Der für alle Aufgaben aus dem Pool im Fach Französisch aggregierte Trennschärfekoeffizient ist als hoch einzustufen (r = 0,84) und unterscheidet sich nicht statistisch signifikant von dem für die landeseigenen Aufgaben ermittelten Koeffizienten (r = 0,90).
- Wie im Fach Englisch wurde auch im Fach Französisch der niedrigste Trennschärfekoeffizient für die Aufgaben zum Kompetenzbereich „Hörverstehen“ gefunden (r = 0,67). Dieser Koeffizient ist signifikant kleiner als der für die Aufgaben zum Kompetenzbereich „Sprachmittlung“ berechnete Kennwert (r = 0,86). Dieser ist wiederum signifikant kleiner als der Trennschärfekoeffizient, der für die Aufgaben zum Kompetenzbereich „Schreiben“ ermittelt wurde (r = 0,95).

Tabelle 15: Aufgabenübergreifende Ergebnisse zur Trennschärfe im Fach Französisch

KB[12]	Anzahl Länder	Trennschärfekoeffizient r	Vertrauensbereich (95 %)
Poolaufgaben			
HV	10	0,67	0,60 / 0,72
SM	14	0,82	0,77 / 0,86
S	8	0,95	0,87 / 0,98
insgesamt	**32**	**0,84***	**0,78 / 0,88**
landeseigene Aufgaben			
insgesamt	**48**	**0,90***	**0,86 / 0,92**

3.3.4 Validität der Aufgaben

Die als Indikatoren zur kriterialen Validität der Aufgaben im Fach Französisch ermittelten Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Tabellen 16 und 17):

- In Bezug auf das Validitätskriterium „Halbjahresnoten“ wurde für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = 0,69 ermittelt. Dieser Koeffizient unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde (r = 0,74). Beide Koeffizienten sind als hoch einzustufen. Deutliche Unterschiede zeigen sich jedoch bei einer nach Kompetenzbereichen differenzierten Betrach-

[12] verwendete Abkürzungen: KB - Kompetenzbereich, HV - Hörverstehen. S - Schreiben, SM - Sprachmittlung

tung. So wurde für die Aufgaben zum Kompetenzbereich „Hörverstehen“ nur ein Validitätskoeffizient mittlerer Höhe (r = 0,46) gefunden. Dieser Kennwert ist signifikant kleiner als die (sich nicht signifikant unterscheidenden) Koeffizienten für die Aufgaben zur „Sprachmittlung“ (r = 0,73) und zum „Schreiben“ (r = 0,83).

- Ein sehr ähnliches Bild zeigt sich in Bezug auf das Validitätskriterium „Klausurnoten“: Der hier für die Aufgaben aus dem Pool berechnete Validitätskoeffizient (r = 0,68) ist ebenfalls als hoch einzustufen und unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben ermittelt wurde (r = 0,73). Wiederum wurde für die Aufgaben zum Kompetenzbereich „Hörverstehen“ (r = 0,52, mittel) ein signifikant kleinerer Validitätskoeffizient gefunden als für die Aufgaben der übrigen beiden Kompetenzbereiche (r = 0,71 bzw. r = 0,84, jeweils hoch).
- Dieses differenzielle Befundmuster für die verschiedenen Kompetenzbereiche weist darauf hin, dass die in den Poolaufgaben erzielten Leistungen in den Kompetenzbereichen „Schreiben“ und „Sprachmittlung“ recht gut die bereits in der Qualifikationsphase erbrachten Ergebnisse abbilden. Deutlich geringer fällt hingegen der Zusammenhang zwischen den Prüfungsleistungen im Kompetenzbereich „Hörverstehen“ einerseits sowie den Halbjahres- und Klausurnoten andererseits aus.

Tabelle 16: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase im Fach Französisch

KB[13]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
HV	10	0,46	0,22 / 0,65
SM	14	0,73	0,66 / 0,79
S	8	0,83	0,78 / 0,87
insgesamt	**32**	**0,69***	**0,62 / 0,76**
landeseigene Aufgaben			
insgesamt	**48**	**0,74***	**0,68 / 0,78**

Tabelle 17: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase im Fach Französisch

KB[13]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
HV	10	0,52	0,47 / 0,58
SM	14	0,71	0,62 / 0,77
S	8	0,84	0,77 / 0,89
insgesamt	**32**	**0,68***	**0,61 / 0,74**
landeseigene Aufgaben			
insgesamt	**48**	**0,73***	**0,67 / 0,77**

[13] verwendete Abkürzungen: KB - Kompetenzbereich, HV - Hörverstehen. S - Schreiben, SM - Sprachmittlung

3.3.5 Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben

Metaanalytisch lassen sich die Ergebnisse für die Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben im Fach Französisch anhand von standardisierten Effektmaßen wie folgt zusammenfassen (vgl. Tabelle 18):

- Insgesamt wurden die Aufgaben aus dem Pool als etwas anspruchsvoller bewertet als die landeseigenen Aufgaben. Der als klein einzustufende Effekt (g = 0,28) ist jedoch statistisch nicht signifikant.
- Bei einer nach Kompetenzbereichen differenzierten Betrachtung zeigt sich, dass die Lehrkräfte den Anspruch der Aufgaben zum „Hörverstehen“ als sehr viel höher eingeschätzt haben als bei den landeseigenen Aufgaben (g = 1.01, starker Effekt). Diese Wahrnehmung korrespondiert mit den Ergebnissen zur empirischen Aufgabenschwierigkeit. Für die Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben zu den übrigen beiden Kompetenzbereichen wurden hingegen nur sehr geringe, statistisch jeweils nicht signifikante Unterschiede zwischen den Aufgaben aus dem Pool und den landeseigenen Aufgaben festgestellt (g = -0,04 bzw. g = -0,13).

Tabelle 18: Aufgabenübergreifende Ergebnisse zum eingeschätzten Anspruch der Aufgaben im Fach Französisch

KB[14]	Anzahl Länder	Mittlere stand. Differenz	Vertrauensbereich (95 %)
HV	9	1,01	0,38 / 1,63
SM	11	-0,13	-0,32 / 0,05
S	4	-0,04	-0,83 / 0,75
insgesamt	**24**	**0,28**	**0,04 / 0,60**

3.3.6 Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

Wie im Fach Englisch wurden auch im Fach Französisch nur geringe, statistisch nicht signifikante Unterschiede in den Einschätzungen der Lehrkräfte zu den EWH der Aufgaben aus dem Pool und der landeseigenen Aufgaben gefunden.

[14] verwendete Abkürzungen: KB - Kompetenzbereich, HV - Hörverstehen. S - Schreiben, SM - Sprachmittlung

3.4 Mathematik

3.4.1 Auswahl der Aufgaben

Anders als in den sprachlichen Fächern haben die Prüflinge in der schriftlichen Abiturprüfung im Fach Mathematik nur in wenigen Ländern die Möglichkeit, zwischen Aufgaben zu wählen. In vier Ländern erfolgt eine Auswahl von Aufgaben durch die Lehrkräfte. Da zur Auswahl von Aufgaben entsprechend keine länderübergreifenden Aussagen getroffen werden können, wurde die Auswahl der Aufgaben für das Fach Mathematik im Rahmen der Evaluation nicht untersucht.

3.4.2 Schwierigkeit der Aufgaben

Im Fach Mathematik wurde für jede Teilaufgabe die empirische Schwierigkeit anhand der Lösungsquote bestimmt, d. h. anhand des Quotienten aus der durchschnittlichen Anzahl der von den Prüflingen bei der betreffenden Teilaufgabe erreichten Bewertungseinheiten (BE) und der Anzahl der dabei maximal erreichbaren BE. Die gefundenen Kennwerte wurden mittels metaanalytischer Methoden zu standardisierten Effektmaßen aggregiert. Die resultierenden Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Tabellen 19 und 20):

- Insgesamt sind die Lösungsquoten bei den Teilaufgaben aus dem Pool statistisch signifikant geringer als bei den landeseigenen Teilaufgaben (g = -0,27, kleiner Effekt). Allerdings findet sich dieses Ergebnismuster nicht konsistent in allen Ländern; in einigen Fällen haben die Prüflinge bei den Teilaufgaben aus dem Pool sogar statistisch signifikant besser abgeschnitten als bei den landeseigenen Teilaufgaben.
- Eine nach Sachgebieten differenzierte Analyse verdeutlicht zudem, dass der für alle Aufgaben ermittelte Effekt offenbar ausschließlich auf das Sachgebiet „Stochastik“ zurückzuführen ist (g = -0,41, kleiner bis mittlerer Effekt). Während hier für die Teilaufgaben aus dem Pool deutlich geringere Lösungsquoten als für die landeseigenen Teilaufgaben festgestellt wurden, fanden sich in den übrigen Sachgebieten („Analysis“ und „Analytische Geometrie/Lineare Algebra“) keine statistisch bedeutsamen Unterschiede (g = 0,03 bzw. g = 0,01, jeweils sehr gering und statistisch nicht signifikant).

Tabelle 19: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Mathematik

	N[15]	Mittlere stand. Differenz	Vertrauensbereich (95 %)
insgesamt	32	-0,27*	-0,42 / -0,12

Tabelle 20: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Sachgebieten im Fach Mathematik

Sachgebiet[16]	*N*[15]	Mittlere stand. Differenz	Vertrauensbereich (95 %)
AG/LA	19	0,01	-0,33 / 0,34
A	24	0,03	-0,26 / 0,31
S	23	-0,41*	-0,58 / -0,24

[15] Die Abiturprüfung wird in vielen Ländern abhängig von den digitalen Hilfsmitteln in verschiedene Varianten durchgeführt. Diese Varianten sind in die Metaanalyse als separate Einheiten (Summe = *N*) eingegangen.

[16] verwendete Abkürzungen: AG/LA - Allgemeine Geometrie/Lineare Algebra, A - Analysis, S - Stochastik

3.4.3 Trennschärfe der Aufgaben

Die mittels metaanalytischer Methoden berechneten Ergebnisse zur Trennschärfe lassen sich wie folgt zusammenfassen (vgl. Tabelle 21):

- Der für alle Teilaufgaben aus dem Pool für das Fach Mathematik aggregierte Trennschärfekoeffizient ist als hoch einzustufen (r = 0,78) und unterscheidet sich nicht statistisch signifikant von dem für die landeseigenen Aufgaben berechneten Koeffizienten (r = 0,79).
- Für die Teilaufgaben aus dem Pool zum Sachgebiet „Stochastik“ wurde ein (immer noch als hoch einzustufender) Trennschärfekoeffizient von r = 0,66 ermittelt. Dieser Koeffizient ist signifikant kleiner als die für die übrigen Sachgebiete berechneten Kennwerte („Analysis“: r = 0,84, „Analytische Geometrie/Lineare Algebra“: r = 0,80).

Tabelle 21: Aufgabenübergreifende Ergebnisse zur Trennschärfe getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Mathematik

Sachgebiet[17]	*N*[18]	Trennschärfekoeffizient *r*	Vertrauensbereich (95 %)
Poolaufgaben			
AG/LA	31	0,80	0,75 / 0,84
A	31	0,84	0,87 / 0,88
S	30	0,66	0,57 / 0,74
insgesamt	**92**	**0,78***	**0,74 / 0,81**
landeseigene Aufgaben			
AG/LA	28	0,79	0,74 / 0,83
A	29	0,85	0,80 / 0,89
S	23	0,70	0,63 / 0,75
insgesamt	**80**	**0,79***	**0,76 / 0,82**

3.4.4 Validität der Aufgaben

Die als Indikatoren zur kriterialen Validität der Aufgaben im Fach Mathematik ermittelten Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Tabellen 22 und 23):

- Der in Bezug auf das Validitätskriterium „Halbjahresnoten“ für die Teilaufgaben aus dem Pool berechnete Validitätskoeffizient ist als hoch einzustufen (r = 0,62) und unterscheidet sich nicht statistisch signifikant von dem Kennwert, der für die landeseigenen Teilaufgaben gefunden wurde (r = 0,64). Der kleinste Validitätskoeffizient wurde für die Teilaufgaben aus dem Pool zum Sachgebiet „Stochastik“ ermittelt (r = 0,52, als mittel einzustufen). Dieser Kennwert ist jeweils signifikant kleiner als die Koeffizienten, die für die Poolaufgaben in den Sachgebieten „Analysis“ (r = 0,67, als hoch einzustufen) und „Analytische Geometrie/Lineare Algebra“ (r = 0,64, als hoch einzustufen) gefunden wurden.
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Teilaufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = 0,59 (als mittel einzustufen) ermittelt. Dieser Koeffizient unterscheidet sich nicht signifikant von dem Kennwert, der

[17] verwendete Abkürzungen: AG/LA - Allgemeine Geometrie/Lineare Algebra, A - Analysis, S - Stochastik

[18] Die Abiturprüfung wird in vielen Ländern abhängig von den digitalen Hilfsmitteln in verschiedene Varianten durchgeführt. Diese Varianten sind in die Metaanalyse als separate Einheiten (Summe = *N*) eingegangen.

für die landeseigenen Aufgaben bestimmt wurde (r = 0,65, als hoch einzustufen). Wiederum wurde der geringste Koeffizient für die Poolaufgaben zum Sachgebiet „Stochastik“ festgestellt (r = 0,51, als mittel einzustufen). Dieser Kennwert ist signifikant kleiner als der für die Aufgaben zur „Analysis“ (r = 0,64, als hoch einzustufen), er unterscheidet sich jedoch nicht signifikant vom Koeffizienten für das Sachgebiet „Analytische Geometrie/Lineare Algebra“ (r = 0,62, als hoch einzustufen).

Tabelle 22: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Mathematik

Sachgebiet[19]	*N*[20]	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
AG/LA	31	0,67	0,61 / 0,72
A	31	0,67	0,62 / 0,71
S	30	0,52	0,44 / 0,59
insgesamt	**92**	**0,62***	**0,59 / 0,66**
landeseigene Aufgaben			
AG/LA	28	0,64	0,60 / 0,68
A	29	0,71	0,65 / 0,75
S	23	0,57	0,51 / 0,63
insgesamt	**80**	**0,65***	**0,62 / 0,68**

Tabelle 23: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Mathematik

Sachgebiet[19]	*N*[20]	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
AG/LA	31	0,62	0,57 / 0,66
A	31	0,64	0,58 / 0,69
S	30	0,51	0,44 / 0,57
insgesamt	**92**	**0,59***	**0,56 / 0,63**
landeseigene Aufgaben			
AG/LA	28	0,63	0,58 / 0,67
A	29	0,71	0,67 / 0,75
S	23	0,60	0,53 / 0,65
insgesamt	**80**	**0,65***	**0,62 / 0,68**

[19] verwendete Abkürzungen: AG/LA - Allgemeine Geometrie/Lineare Algebra, A - Analysis, S - Stochastik

[20] Die Abiturprüfung wird in vielen Ländern abhängig von den digitalen Hilfsmitteln in verschiedene Varianten durchgeführt. Diese Varianten sind in die Metaanalyse als separate Einheiten (Summe = *N*) eingegangen.

- Dieses differenzielle Befundmuster für die verschiedenen Sachgebiete weist darauf hin, dass die in den Poolaufgaben erzielten Leistungen in den Sachgebieten „Analysis“ und „Analytische Geometrie/Lineare Algebra“ recht gut die bereits in der Qualifikationsphase erbrachten Ergebnisse abbilden. Weniger hoch fällt der Zusammenhang zwischen den Prüfungsleistungen sowie den Halbjahres- und Klausurnoten hingegen im Sachgebiet „Stochastik“ aus.

3.4.5 Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben

Für das Fach Mathematik wurden die Einschätzungen zum Anspruch der Aufgaben nicht bezogen auf Teilaufgaben, sondern bezogen auf vollständige Aufgaben bzw. Prüfungsteile der jeweils landesspezifischen Abiturprüfung erhoben. Eine Interpretation der Ergebnisse in Bezug auf die aus dem Pool eingesetzten Teilaufgaben ist damit nur eingeschränkt möglich. Eine Aufgabe bzw. ein Prüfungsteil wurde dabei immer dann als aus dem Pool entnommen betrachtet, wenn die darin enthaltenen Teilaufgaben im Umfang von mehr als der Hälfte der maximal erreichbaren Bewertungseinheiten aus dem Pool stammen. Metaanalytisch lassen sich die dazu ermittelten Ergebnisse wie folgt zusammenfassen (vgl. Tabelle 24):

Im Vergleich zu den landeseigenen Aufgaben haben die Lehrkräfte den Anspruch der Poolaufgaben im Fach Mathematik insgesamt als höher eingeschätzt (g = 0,16). Dieser Unterschied ist jedoch sehr gering und außerdem nicht statistisch signifikant.

Tabelle 24: Aufgabenübergreifende Ergebnisse zum eingeschätzten Anspruch der Aufgaben im Fach Mathematik

	N[21]	Mittlere stand. Differenz	Vertrauensbereich (95 %)
insgesamt	30	0,16	-0,03 / 0,34

3.4.6 Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

Anders als für die sprachlichen Fächer wurde der Nutzen der EWH im Fach Mathematik nicht aufgabenspezifisch erfasst. Zum einen wurde der Informationsgehalt einer aufgabenspezifischen Abfrage von Einschätzungen zum EWH für das Fach Mathematik als eher gering beurteilt. Zum anderen wäre eine differenzierte Erfassung für die Lehrkräfte sehr aufwändig gewesen. Es wurde jedoch aufgabenübergreifend erhoben, wie hilfreich die EWH insgesamt für die Bewertung der Aufgaben von den Lehrkräften eingeschätzt wurden. Da die Auswertung dieser Angaben nur landesspezifisch erfolgen kann und landesspezifische Ergebnisse nicht Gegenstand dieses Berichts sind, wird an dieser Stelle auf eine Darstellung verzichtet.

[21] Die Abiturprüfung wird in vielen Ländern abhängig von den digitalen Hilfsmitteln in verschiedene Varianten durchgeführt. Diese Varianten sind in die Metaanalyse als separate Einheiten (Summe = N) eingegangen.

4 Literatur

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin: Springer.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik: Mit Hinweisen zur Intervention* (2. Aufl.). Göttingen: Hogrefe.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. and Rothstein, H. R. 2009. *Introduction to meta-analysis*, Chichester, UK: Wiley.