Institut zur Qualitätsentwicklung
im Bildungswesen

Gemeinsame Abituraufgabenpools der Länder

Evaluation von Aufgaben der Pools für das Prüfungsjahr 2019

Ergebnisse zur Bewährung der Aufgaben

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

Inhalt

1 Kurzzusammenfassung

Der vorliegende Bericht stellt die Ergebnisse der Evaluation zum Einsatz von Abiturprüfungsaufgaben der Pools für das Prüfungsjahr 2019 dar. Anders als bei der Evaluation für das Prüfungsjahr 2017 wurde dabei nur für die Fächer Deutsch, Englisch und Französisch untersucht, wie sich die Aufgaben der Abituraufgabenpools für das Prüfungsjahr 2019 bewährt haben. Insgesamt zeigten sich in den sechs betrachteten Evaluationsbereichen zwischen den Aufgaben aus den Pools und den landeseigenen Aufgaben nur noch sehr vereinzelt signifikante Unterschiede, d. h. in deutlich geringerer Anzahl als im Prüfungsjahr 2017. Dies könnte ein erster Hinweis darauf sein, dass die mit dem Projekt intendierten normierenden Wirkungen der Poolaufgaben auf die anderen Prüfungsaufgaben der Länder tatsächlich auftreten. Fachspezifisch sind weiterhin einzelne Auffälligkeiten zu verzeichnen, die wichtige Hinweise für die ländergemeinsame Entwicklung von Abiturprüfungsaufgaben geben:

- Im Fach Deutsch wurden die aus dem Abituraufgabenpool stammenden Aufgaben von den Schülerinnen und Schülern relativ häufig ausgewählt, insbesondere die Aufgaben „Herzschmerz“ (Interpretation literarischer Texte) und „Freigabe“ (Materialgestütztes Verfassen informierender Texte). Die Notenpunkte, die bei der Bearbeitung von Aufgaben der Art „Interpretation literarischer Texte“ aus dem Pool im Mittel erzielt wurden, fielen höher aus als bei den zum Vergleich herangezogenen landeseigenen Aufgaben derselben Aufgabenart. Dieser Befund könnte (wie im Prüfungsjahr 2017) dadurch bedingt sein, dass Aufgaben der Art „Interpretation literarischer Texte“ häufiger von Prüflingen mit besseren Vorleistungen in der Qualifikationsphase gewählt wurden. Der Effekt dürfte allerdings auch auf eine einzelne Aufgabe zurückgehen, die von den befragten Lehrkräften im Durchschnitt als weniger anspruchsvoll als die jeweiligen landeseigenen Aufgaben eingeschätzt wurde.
- Im Fach Englisch zeigten sich (wie bereits im Prüfungsjahr 2017) geringe Auffälligkeiten bei den Poolaufgaben zum Kompetenzbereich „Hörverstehen“. Die hier erzielten Ergebnisse unterschieden sich zwar nicht von den bei der Bearbeitung landeseigener Aufgaben erzielten Resultaten, wiesen im Vergleich zu den Ergebnissen bei der Bearbeitung von Aufgaben anderer Kompetenzbereiche allerdings einen deutlich geringeren Zusammenhang mit den Vorleistungen der Prüflinge auf. Allerdings fanden sich zwischen den Aufgaben aus dem Pool und den landeseigenen Aufgaben insgesamt keine Unterschiede in den Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben aller Kompetenzbereiche. Im Prüfungsjahr 2017 wurden die Aufgaben des Pools zum Kompetenzbereich „Hörverstehen“ im Vergleich zu den landeseigenen Aufgaben noch deutlich anspruchsvoller eingeschätzt.
- Ähnlich wie im Fach Englisch zeigte sich auch im Fach Französisch für den Kompetenzbereich „Hörverstehen“ im Vergleich zu den anderen Kompetenzbereichen ein geringerer Zusammenhang zwischen den bei der Bearbeitung der Poolaufgaben erzielten Ergebnissen und den Vorleistungen der Prüflinge. Anders als im Prüfungsjahr 2017 fanden sich jedoch bei der empirischen Aufgabenschwierigkeit und beim eingeschätzten Anspruch keine signifikanten Unterschiede zwischen den Poolaufgaben zum „Hörverstehen“ und den landeseigenen Aufgaben insgesamt.

2 Methodisches Vorgehen

Mit Beschluss der „Konzeption zur Implementation der Bildungsstandards für die Allgemeine Hochschulreife“ hat die KMK am 10.10.2013 das IQB beauftragt, die Entwicklung und Nutzung der gemeinsamen Abituraufgabenpools der Länder auch wissenschaftlich zu evaluieren. Hauptziel dieser Evaluation ist gemäß der genannten Konzeption, „[…] Evidenz dafür zu erbringen, dass die mit de[n] Aufgabenpool[s] angestrebten Funktionen erreicht werden“. Vor diesem Hintergrund wird der Einsatz von Aufgaben aus den Pools seit dem Prüfungsjahr 2017 von einer formativen Evaluation begleitet, die Anhaltspunkte dafür gewinnen soll, ob sich diese Aufgaben bewähren.

2.1 Kernpunkte des Evaluationskonzepts

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

- Auswahl der Aufgaben (Wie häufig werden – sofern Wahlmöglichkeiten bestehen – die Aufgaben aus den Pools im Vergleich zu den landeseigenen Aufgaben von den Prüflingen gewählt?)
- Empirische Schwierigkeit der Aufgaben (Unterscheidet sich die empirische Schwierigkeit der Aufgaben aus den Pools von der empirischen Schwierigkeit der landeseigenen Aufgaben?)
- Trennschärfe der Aufgaben (Unterscheidet sich die Trennschärfe[1] der Aufgaben aus den Pools von der Trennschärfe landeseigener Aufgaben?)
- Kriteriale Validität der Aufgaben (Gibt es einen Zusammenhang zwischen den Ergebnissen, die bei den Aufgaben aus den Pools erzielt wurden, und den Vorleistungen[2] der Prüflinge?)
- Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben (Wie verhält sich die Einschätzung zum Anspruch der Aufgaben aus den Pools im Vergleich zu landeseigenen Aufgaben?)
- Einschätzungen der Lehrkräfte zu den Erwartungshorizonten der Aufgaben (Wie verhält sich die Einschätzung zu den Erwartungshorizonten der Aufgaben aus den Pools zur Einschätzung zu den Erwartungshorizonten der landeseigenen Aufgaben?)

Im Rahmen der Evaluation erfolgte zudem eine vertiefte Betrachtung ausgewählter Aufgabenarten bzw. Kompetenzbereiche. Hierbei wurden die an der Evaluation teilnehmenden Lehrkräfte im Fach Deutsch insbesondere zu den Aufgaben der Arten „Materialgestütztes Verfassen argumentierender Texte“ und „Materialgestütztes Verfassen informierender Texte“ um zusätzliche Einschätzungen gebeten. In den Fächern Englisch und Französisch lag der Schwerpunkt der zusätzlichen Lehrkräftebefragung auf den Aufgaben der Kompetenzbereiche „Schreiben“ und „Hörverstehen“.

[1] 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).

[2] Als Indikatoren für die Vorleistungen der Prüflinge werden die Halbjahresnoten der Qualifikationsphase und die in der Qualifikationsphase erzielten Klausurergebnisse betrachtet.

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 2019, die von den Ländern entnommen wurden, können auf den Internetseiten des IQB unter https://www.iqb.hu-berlin.de/abitur/evaluation eingesehen werden. [3]

An der Evaluation zur Bewährung der Aufgaben aus den Pools für das Prüfungsjahr 2019 nahmen alle 16 Länder teil. Ein für die Datenerhebung vom IQB bereitgestelltes elektronisches Eingabeinstrument wurde von 14 Ländern genutzt. Die Länder Brandenburg und Nordrhein-Westfalen übermittelten Daten aus ihren landeseigenen Erhebungen. Im Land Bremen wurde ein Teil der für die Evaluation benötigten Daten mithilfe des Eingabeinstruments des IQB, der andere Teil im Rahmen einer eigenen Erhebung erfasst und an das IQB übermittelt.

In den Ländern Brandenburg und Nordrhein-Westfalen waren die Schulen, an denen die für die Evaluation benötigten Daten erhoben wurden, vom jeweils zuständigen Landesinstitut ausgewählt worden. In den anderen Ländern wurde die jeweilige Stichprobe in Abstimmung mit den jeweiligen Ansprechpartnerinnen und -partnern der Länder durch das IQB gezogen. Dazu wurden die Länder gebeten, dem IQB anonymisierte Listen aller in Frage kommenden Schulen zur Verfügung zu stellen. Auf dieser Grundlage erfolgte per Zufallsauswahl die Ziehung der Schulstichproben, die pro Land 20 Schulen umfassen.[4] Hierbei wurde für jedes Land darauf geachtet, dass das Verhältnis zwischen Gymnasien und anderen allgemeinbildenden Schulen mit einer gymnasialen Oberstufe (z. B. Gesamtschulen) dem entsprechenden Verhältnis in der Grundgesamtheit entspricht.

An jeder Schule wurde für jedes der drei Fächer und für jedes Anforderungsniveau ein Kurs in die Datenerhebung einbezogen, sofern in der entsprechenden Abiturprüfung Aufgaben aus den Pools eingesetzt wurden und Prüflinge eine Prüfung abgelegt haben.[5] Damit waren maximal 6 Kurse pro Schule an der Datenerhebung beteiligt.[6] Die Auswahl der Kurse erfolgte jeweils durch Verantwortliche der Schulleitung. Um Lehrkräfte großer Kurse zu entlasten, wurde die Anzahl der Prüflinge pro Kurs, deren Prüfungsergebnisse eingegeben werden sollten, auf 20 beschränkt.[7,8]

2.3 Überblick zur Stichprobe

Angaben zur Anzahl der in die Evaluation einbezogenen Prüfungsarbeiten können Tabelle 1 entnommen werden.

---

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

[4] Auf Grund der geringen Anzahl der für die Evaluation zur Verfügung stehenden Schulen umfasste die Schulstichprobe in den Ländern Bremen und Hamburg nur jeweils 10 Schulen.

[5] In einigen Ländern wurde bereits vorab auf eine Datenerhebung im Fach Französisch verzichtet, da die zu erwartenden Fallzahlen keine statistisch belastbaren Analysen erlaubt hätten.

[6] Um trotz der geringen Schulstichprobe hinreichend viele Informationen zur Bestimmung der mit Blick auf die Bewährung der Aufgaben betrachteten Indikatoren zu haben, wurde diese Beschränkung für das Land Hamburg aufgehoben.

[7] Die Auswahl der Schülerinnen und Schüler erfolgte durch die zuständigen Lehrkräfte. Diese wurden hierfür wie folgt instruiert: Die Auswahl soll „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)“.

[8] Aufgrund der kleineren Schulstichprobe wurde diese Begrenzung für das Land Bremen aufgehoben.

Tabelle 1: Stichprobengröße

	Deutsch	Englisch	Französisch
Anzahl der Prüfungsarbeiten	6903	8589	1346
… auf erhöhtem Niveau	4995	6532	1302
… auf grundlegendem Niveau	1908	2057	44

Auffällig ist die sehr kleine Analysestichprobe für das grundlegende Niveau im Fach Französisch. Aufgrund der geringen Stichprobengröße wird hier auf eine Darstellung von Ergebnissen zu den Evaluationsbereichen „Auswahl der Aufgaben“ und „empirische Schwierigkeit der Aufgaben“ verzichtet.

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).[9] 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.

[9] 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 drei 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

In der schriftlichen Abiturprüfung im Fach Deutsch können die Prüflinge in allen Ländern aus mindestens zwei Alternativen eine Abiturprüfungsaufgabe auswählen.[10] Das Ergebnis dieser Aufgabenauswahl (im Folgenden auch als „Attraktivität der Aufgaben“ bezeichnet) ist Gegenstand der Evaluation, d. h., es wird geprüft, wie häufig die aus dem Pool stammenden Prüfungsaufgaben im Vergleich zu den landeseigenen Aufgaben gewählt wurden. Insgesamt lassen sich diese Ergebnisse wie folgt zusammenfassen:

- Viele der aus dem Pool stammenden Aufgaben wurden von den Prüflingen relativ häufig ausgewählt.
- Besonders hohe Attraktivitätswerte lassen sich in drei Ländern für die Aufgabe „Herzschmerz“ (Interpretation literarischer Texte) sowie für die Aufgabe „Freigabe“ (Materialgestütztes Verfassen informierender Texte) feststellen.
- In wenigen Fällen finden sich geringe Attraktivitätswerte: für die Aufgabe „Okay“ (Analyse pragmatischer Texte), für die Aufgabe „Schwefel“ (Materialgestütztes Verfassen informierender Texte) sowie in jeweils einem Land für die Aufgaben „Konserve“ und „Herzschmerz“ (jeweils Interpretation literarischer Texte).

In einer Metaanalyse wurde aufgabenübergreifend der Zusammenhang zwischen der Auswahl der Aufgaben und bestimmten Hintergrundmerkmalen der Prüflinge betrachtet. Unter anderem wurde untersucht, ob die Vorleistungen der Prüflinge Einfluss auf die Auswahl der unterschiedlichen Aufgabenarten haben. Die Ergebnisse können Tabelle 3 entnommen werden uns lassen sich wie folgt zusammenfassen:

- Anders als bei der letztmaligen Evaluation der Bewährung der Aufgaben aus den Pools für das Prüfungsjahr 2017 wurden für das Prüfungsjahr 2019 zwischen denjenigen Prüflingen, die eine aus dem Pool stammende Aufgabe gewählt haben, und denjenigen, die eine landeseigene Aufgabe bearbeitet haben, keine signifikanten Unterschiede in den Vorleistungen gefunden (Differenz: -0.09 NP, nicht statistisch signifikant).
- Analog zum Prüfungsjahr 2017 wurde für das Prüfungsjahr 2019 festgestellt, dass Aufgaben der Art „Interpretation literarischer Texte“ häufiger von Prüflingen gewählt wurden, die im Vergleich zu anderen Schülerinnen und Schülern signifikant bessere Vorleistungen aufweisen (Differenz: +0.37 NP, statistisch signifikant).

[10] In einigen Fällen nimmt die Schule oder die zuständige Lehrkraft zusätzlich eine Vorauswahl vor.

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

Aufgabenart[11]	Anzahl Länder	Differenz (in NP)	Vertrauensbereich (95 %)
AP	6	-0.51	[-1.10; 0.08]
EP	3	-0.66	[-3.16; 1.83]
IL	14	0.37*	[0.03; 0.71]
MA	3	-0.47	[-1.23; 0.28]
MI	2	-0.13	[-11.26;10.98]
insgesamt	**28**	**-0.09**	**[-0.38; 0.20]**

3.1.2 Schwierigkeit der Aufgaben

Als Indikatoren für die empirische Schwierigkeit der aus dem Pool für das Fach Deutsch eingesetzten Aufgaben wurden jeweils die arithmetischen Mittelwerte der von den Prüflingen erzielten Notenpunkte berechnet. Zudem wurde untersucht, ob sich die eingesetzten Aufgaben aus dem Pool hinsichtlich der Schwierigkeit von den landeseigenen Aufgaben unterscheiden. Die Ergebnisse können Tabelle 4 entnommen werden uns lassen sich wie folgt zusammenfassen:

- Die Ergebnisse der Metaanalyse lassen darauf schließen, dass zwischen den Poolaufgaben einerseits und den landeseigenen Aufgaben andererseits hinsichtlich der von den Prüflingen erzielten Ergebnisse kein signifikanter Unterschied besteht (Differenz: -0.12 NP, nicht statistisch signifikant).
- Für die Aufgaben der Art „Interpretation literarischer Texte“ fallen die Ergebnisse der Prüflinge allerdings für die Poolaufgaben signifikant besser aus als für die zum Vergleich herangezogenen landeseigenen Aufgaben (Differenz: +0.47 NP, statistisch signifikant). Wie im Abschnitt 3.1.1 skizziert, könnte dieser Befund zum einen dadurch bedingt sein, dass die Aufgaben der Art „Interpretation literarischer Texte“ häufiger von leistungsstärkeren Schülerinnen und Schülern gewählt wurden. Zum anderen könnte der festgestellte Effekt darauf zurückgeführt werden, dass die Interpretationsaufgabe „Herzschmerz“ in insgesamt sieben Ländern eingesetzt wurde und dementsprechend in den Metaanalysen ein hohes Gewicht erhält. In fünf der sieben Länder haben die Prüflinge bei dieser Aufgabe signifikant besser abgeschnitten als bei den jeweiligen landeseigenen Aufgaben. Von den im Rahmen der Evaluation befragten Lehrkräften wurde die Aufgabe gleichzeitig als signifikant weniger anspruchsvoll beurteilt als die jeweiligen landeseigenen Aufgaben.
- Für alle übrigen Aufgabenarten wurden im Rahmen der Metanalysen zur empirischen Schwierigkeit der Aufgaben keine signifikanten Unterschiede zwischen den Poolaufgaben und den landeseigenen Aufgaben ermittelt.

[11] 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[12]	Anzahl Länder	Differenz (in NP)	Vertrauensbereich (95 %)
AP	6	-0.48	[-1.08; 0.12]
EP	3	-1.21	[-3.97; 1.56]
IL	14	0.47*	[0.00; 0.93]
MA	3	-0.23	[-1.08; 0.62]
MI	2	-0.87	[-6.73; 5,00]
insgesamt	**28**	**-0.12**	**[-0.48; 0.24]**

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 Kriteriale 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 metaanalytisch für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = .76 ermittelt. Anders als bei der Evaluation für das Prüfungsjahr 2017 unterscheidet sich dieser Koeffizient nicht signifikant von dem Koeffizienten, der für die landeseigenen Aufgaben bestimmt wurde (r = .75). Beide Koeffizienten sind als hoch einzustufen. Auch die für die einzelnen Aufgabenarten berechneten Validitätskoeffizienten (zwischen r = .65 und r = .79) sind bei den Poolaufgaben als hoch zu bewerten.
- In Bezug auf das Validitätskriterium „Klausurnoten im Fach Deutsch“ wurde metaanalytisch für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = .78 ermittelt. Dieser Koeffizient ist in seiner Höhe identisch mit dem Koeffizienten, der für die landeseigenen Aufgaben bestimmt wurde (r = .78). Beide Koeffizienten sind als hoch einzustufen. Auch die für die einzelnen Aufgabenarten berechneten Validitätskoeffizienten (zwischen r = .69 bis r = .81) sind als hoch zu bewerten.

[12] 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[13]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
AP	6	0.75	[0.71; 0.78]
EP	3	0.79	[0.78; 0.81]
IL	14	0.70	[0.43; 0.86]
MA	3	0.65	[-0.98; 1.00]
MI	2	0.69	[0.39; 0.86]
insgesamt	**28**	**0.76**	**[0.74; 0.78]**
landeseigene Aufgaben			
insgesamt	**49**	**0.75**	**[0.73; 0.78]**

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	6	0.78	[0.70; 0.84]
EP	3	0.81	[0.79; 0.84]
IL	14	0.71	[0.51; 0.84]
MA	3	0.69	[-0.62; 0.99]
MI	2	0.72	[0.30; 0.91]
insgesamt	**28**	**0.78**	**[0.75; 0.80]**
landeseigene Aufgaben			
insgesamt	**49**	**0.78**	**[0.76; 0.80]**

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 anspruchsvoller eingeschätzt als die landeseigenen Aufgaben. Der Unterschied ist allerdings nicht signifikant (g = .15, nicht statistisch signifikant).

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

- Auch bei einer nach Aufgabenarten differenzierten Betrachtung zeigen die Einschätzungen der Lehrkräfte zum Anspruch keine signifikanten Unterschiede zwischen den Poolaufgaben und den landeseigenen Aufgaben.
- Wie bereits unter 3.1.2 beschrieben, fällt insbesondere die Interpretationsaufgabe „Herzschmerz“ auf, die in fast allen Ländern, die diese Aufgabe aus dem Pool entnommen haben, von den Lehrkräften signifikant als weniger anspruchsvoll bewertet wurde als die jeweiligen landeseigenen Aufgaben.

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

Aufgabenart[14]	**Anzahl Länder**	**Stand. Differenz (*g*)**	**Vertrauensbereich (95 %)**
AP	6	0.46	[-0.17; 1.08]
EP	3	-0.18	[-2.35; 1.99]
IL	14	-0.04	[-0.41; 0.33]
MA	3	0.49	[-0.41; 1.40]
MI	2	0.47	[-7.38; 8.31]
insgesamt	28	0.15	[-0.11; 0.40]

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 Erwartungshorizonte der Aufgaben aus dem Pool als ähnlich hoch beurteilt wie die Erwartungshorizonte der landeseigenen Aufgaben (g = .03, nicht statistisch signifikant).
- Auch bei einer nach Aufgabenarten differenzierten Betrachtung zeigen die Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte keine statistisch signifikanten Unterschiede zwischen den Poolaufgaben und den landeseigenen Aufgaben.
- Bei einer aufgabenspezifischen Betrachtung fällt auf, dass der Erwartungshorizont zur Aufgabe „Herzschmerz“ (Interpretation literarischer Texte) in allen Ländern, in denen die Aufgabe eingesetzt wurde, signifikant positiver beurteilt wurde als die Erwartungshorizonte der jeweiligen landeseigenen Aufgaben. Demgegenüber wurden die Erwartungshorizonte der Aufgaben „Zeh“ (Interpretation literarischer Texte) und „Ehebruch“ (Analyse pragmatischer Texte) von den Lehrkräften jeweils eines Landes als signifikant weniger nützlich ein-geschätzt als die Erwartungshorizonte der jeweiligen landeseigenen Aufgaben.

[14] 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 8: Aufgabenübergreifende Ergebnisse zum eingeschätzten Nutzen der EWH der Aufgaben im Fach Deutsch

Aufgabenart[7]	Anzahl Länder	Stand. Differenz (*g*)	Vertrauensbereich (95 %)
AP	6	-0.09	[-0.54; 0.36]
EP	3	-0.22	[-2.36; 1.92]
IL	14	0.23	[-0.02; 0.47]
MA	3	-0.29	[-1.28; 0.70]
MI	2	-0.39	[-1.35; 0.57]
insgesamt	28	0.03	[-0.15; 0.20]

3.1.7 Weitere Einschätzungen und Kommentare der Lehrkräfte zu den Aufgaben

Die zentralen Ergebnisse für die weiteren im Rahmen der Evaluation erhobenen Lehrkräfteurteile zu den Aufgaben der Pools lassen sich wie folgt zusammenfassen:

- In den meisten Ländern und für die meisten Aufgaben haben die befragten Lehrkräfte angegeben, dass die Schülerinnen und Schüler in der Qualifikationsphase intensiv auf die Anforderungen der jeweiligen Aufgabenart vorbereitet worden seien. Allerdings gibt es hierbei auch einige wenige Ausnahmen: So lässt sich in einem Land anhand der Einschätzungen zu den Aufgaben der Art „Interpretation literarischer Texte“ „Herzschmerz“ (erhöhtes Niveau) und „Konserve“ (grundlegendes Niveau) erkennen, dass die Schülerinnen und Schüler in diesem Land offenbar insgesamt weniger intensiv auf die Anforderungen einer Gedichtinterpretation vorbereitet worden sind. Tatsächlich wurden diese beiden Aufgaben in dem betreffenden Land auch nur von sehr wenigen Prüflingen gewählt. Eine etwas weniger intensive Vorbereitung wurde auch für die Aufgabe „Schwefel“ der Art „materialgestütztes Verfassen informierender Texte“ angegeben, die von einem Land auf erhöhtem Niveau eingesetzt und ebenfalls nur von wenigen Prüflingen gewählt wurde.
- In Bezug auf die Aufgaben der Arten „materialgestütztes Verfassen argumentierender Texte“ und „materialgestütztes Verfassen informierender Texte“ wird häufig problematisiert, dass der zu erfassende Zieltext keine lebensweltlich relevante Textsorte abbilde und die Situierung der Aufgabenstellung wenig realistisch sei. Auch die hohe Anzahl der Materialien wird häufig kritisch gesehen.
- Die Erwartungshorizonte zu den Aufgaben des Pools werden häufig als etwas zu detailliert und zu umfangreich bewertet. Oft findet sich auch die Anmerkung, dass die im Erwartungshorizont formulierten Anforderungen unrealistisch hoch seien.

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 aus organisatorischen Gründen nicht möglich, Wahlmöglichkeiten anzubieten, da die Schülerinnen und Schüler die Aufgaben gleichzeitig bearbeiten. Insgesamt zeigen die Ergebnisse, dass die Aufgaben aus dem Pool im Vergleich zu den landeseigenen Aufgaben ähnlich häufig bzw. tendenziell etwas seltener gewählt wurden.

Aufgrund der eingeschränkten Wahlmöglichkeiten im Fach Englisch ist es für die Prüfungen vieler 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):

- Wie bei der Evaluation für das Prüfungsjahr 2017 wurde auch für das Prüfungsjahr 2019 kein signifikanter Unterschied zwischen den bei den Aufgaben aus dem Pool erreichten Ergebnissen und den bei den landeseigenen Aufgaben erreichten Ergebnissen festgestellt (g = -.02, nicht statistisch signifikant).
- In der Evaluation für das Prüfungsjahr 2017 war ermittelt worden, dass die Prüflinge bei den Aufgaben zum Hörverstehen signifikant besser abgeschnitten hatten als bei den landeseigenen Aufgaben. Für das Prüfungsjahr 2019 findet sich ein solcher Effekt nicht (g = .01, nicht statistisch signifikant). Leider kann anhand der zur Verfügung stehenden Daten nicht bestimmt werden, wodurch dieses Befundmuster bedingt ist. Denkbar erscheint, dass die Aufgaben zum Hörverstehen etwas leichter waren als im Prüfungsjahr 2017. Möglich ist aber auch, dass die landeseigenen Aufgaben etwas schwerer ausgefallen sind als im Prüfungsjahr 2017.
- Analog zum Prüfungsjahr 2017 wurden im Kompetenzbereich „Sprachmittlung“ keine statistisch signifikanten Unterschiede zwischen den in Poolaufgaben und den in landeseigenen Aufgaben erzielten Ergebnissen gefunden (g = .07, nicht statistisch signifikant). Für den Kompetenzbereich „Schreiben“ lässt sich zwar für die Poolaufgaben eine etwas höhere empirische Aufgabenschwierigkeit feststellen. Dieser Effekt ist allerdings gering und statistisch nicht signifikant (g = -.22, nicht statistisch signifikant).

Tabelle 9: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Englisch

KB[15]	Anzahl Länder	Stand. Differenz (*g*)	Vertrauensbereich (95 %)
HV	10	0.01	[-0.40; 0.43]
SM	11	-0.22	[-0.40; -0.04]
S	19	0.07	[-0.08; 0.22]
insgesamt	**40**	**-0.02**	**[-0.14; 0.11]**

3.2.3 Trennschärfe der Aufgaben

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 = .84) und unterscheidet sich nicht signifikant von dem für die landeseigenen Aufgaben berechneten Koeffizienten (r = .90).
- Für die Poolaufgaben zum „Hörverstehen“ wurde ein aggregierter Trennschärfekoeffizient von r = .75 gefunden. Dieser Koeffizient ist signifikant kleiner als der Kennwert für die Poolaufgaben zum „Schreiben“ (r = .93), unterscheidet sich jedoch nicht signifikant vom Koeffizienten für die Poolaufgaben zur „Sprachmittlung“ (r = .82). Insgesamt ähnelt dieses nach Kompetenzbereichen differenzierte Befundmuster stark den Ergebnissen, die im Rahmen der Evaluation für das Prüfungsjahr 2017 ermittelt wurden.

Tabelle 10: Aufgabenübergreifende Ergebnisse zur Trennschärfe im Fach Englisch

KB[8]	Anzahl Länder	Trennschärfekoeffizient *r*	Vertrauensbereich (95 %)
Poolaufgaben			
HV	18	0.82	[0.77; 0.87]
SM	13	0.93	[0.87; 0.96]
S	13	0.75	[0.69; 0.80]
insgesamt	**44**	**0.84**	**[0.80; 0.87]**
landeseigene Aufgaben			
insgesamt	**57**	**0.90**	**[0.87; 0.92]**

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

3.2.4 Kriteriale 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 = .70 ermittelt. Dieser unterscheidet sich nicht signifikant von dem Koeffizienten, der für die landeseigenen Aufgaben bestimmt wurde (r = .75). Beide Koeffizienten sind als hoch einzustufen. Bei einer nach Kompetenzbereichen differenzierten Analyse findet sich der höchste Validitätskoeffizient für die Aufgaben zum „Schreiben“ (r = .77). Auch für die Sprachmittlung ist der Koeffizient mit r = .73 recht hoch ausgeprägt. Dieser Wert ist signifikant höher als der für die Aufgaben zum „Hörverstehen“ ermittelte Koeffizient (r = .61).
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Aufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = .74 ermittelt. Dieser unterscheidet sich nicht signifikant von dem Koeffizienten, der für die landeseigenen Aufgaben bestimmt wurde (r = .80). Während die für die Poolaufgaben zum „Schreiben“ (r = .80) und zur „Sprachmittlung“ (r = .76) gefundenen Validitätskoeffizienten nahezu identisch sind, fällt der für die Poolaufgaben zum „Hörverstehen“ berechnete Koeffizient (r = .65) wiederum signifikant geringer aus.
- Insgesamt weisen die ermittelten Validitätskoeffizienten darauf hin, dass die in den Poolaufgaben aller Kompetenzbereiche erzielten Leistungen recht gut die in der Qualifikationsphase erzielten Ergebnisse abbilden. Damit ähnelt das für den Evaluationsbereich der kriterialen Validität festgestellte Befundmuster den Ergebnissen, die im Rahmen der Evaluation für das Prüfungsjahr 2017 ermittelt wurden.

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

KB[16]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
HV	18	0.73	[0.70; 0.75]
SM	13	0.77	[0.68; 0.83]
S	13	0.61	[0.55; 0.65]
insgesamt	**44**	**0.70**	**[0.67; 0.73]**
landeseigene Aufgaben			
insgesamt	**57**	**0.75**	**[0.72; 0.78]**

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

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	´0	0.76	[0.74; 0.79]
SM	13	0.80	[0.73; 0.85]
S	13	0.65	[0.62; 0.69]
insgesamt	**44**	**0.74**	**[0.71; 0.77]**
landeseigene Aufgaben			
insgesamt	**57**	**0.80**	**[0.76; 0.82]**

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 nicht signifikant und zudem als gering einzustufen (g = .14, nicht statistisch signifikant).
- Bei einer nach Kompetenzbereichen differenzierten Betrachtung fällt auf, dass insbesondere die Poolaufgaben zum Hörverstehen und zum Schreiben als etwas anspruchsvoller als die landeseigenen Aufgaben bewertet wurden. Allerdings sind auch diese Unterschiede jeweils als gering und nicht signifikant einzustufen (g = .23 bzw. g = .26, jeweils nicht statistisch signifikant). Dieses Befundmuster unterscheidet sich etwas von dem Ergebnis der Evaluation für das Prüfungsjahr 2017, bei der die Aufgaben zum Kompetenzbereich „Hörverstehen“ im Vergleich zu den landeseigenen Aufgaben als deutlich anspruchsvoller bewertet wurden.

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

KB[17]	Anzahl Länder	Mittlere stand. Differenz	Vertrauensbereich (95 %)
HV	12	0.23	[-0.13; 0.59]
SM	8	0.26	[-0.09; 0.60]
S	15	0.00	[-0.22; 0.23]
insgesamt	**35**	**0.14**	**[-0.02; 0.30]**

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

3.2.6 Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

Im Hinblick auf den Nutzen der Erwartungshorizonte finden sich sowohl in der Gesamtschau als auch bei einer nach Kompetenzbereichen differenzierten Betrachtung nur geringe Unterschiede (g = .03, nicht statistisch signifikant) in den Einschätzungen der Lehrkräfte zwischen den Aufgaben aus dem Pool und den landeseigenen Aufgaben. Auch dieses Befundmuster ähnelt den Ergebnissen, die im Rahmen der Evaluation für das Prüfungsjahr 2017 ermittelt wurden.

3.2.7 Weitere Einschätzungen und Kommentare der Lehrkräfte zu den Aufgaben

Die zentralen Ergebnisse für die weiteren im Rahmen der Evaluation erhobenen Lehrkräfteurteile zu den Aufgaben der Pools lassen sich wie folgt zusammenfassen:

- In den meisten Ländern und für die meisten Aufgaben haben die befragten Lehrkräften angegeben, dass die Schülerinnen und Schüler in der Qualifikationsphase recht intensiv auf die jeweiligen Anforderungen der betreffenden Aufgabenart vorbereitet worden seien. Ein etwas geringerer Wert kann (allerdings nur in einem Land) einzig für die Aufgabe „Newspapers“ (Kompetenzbereich „Schreiben“) festgestellt werden.
- Die zusätzlichen Einschätzungen der Lehrkräfte zum Kompetenzbereich „Hörverstehen“ zeigen für die aus dem Pool entnommenen Aufgaben ein insgesamt eher positives Bild. Allerdings gibt es hierbei auch einige wenige Ausnahmen: Für die Kombination der beiden Aufgaben „Alan“ und „Money“ fällt in einem Land die von den Lehrkräften beurteilte Angemessenheit der Einlese- und Bearbeitungszeit verhältnismäßig gering aus. Bei der Aufgabe „Acting“ werden von einigen Lehrkräften die Sprechgeschwindigkeit, die Komplexität der Sprache und die Häufigkeit des Sprecherwechsels problematisiert.
- In den Kommentaren der Lehrkräfte zu den Aufgaben des Kompetenzbereichs „Hörverstehen“ wird bei einigen Aufgaben (z. B. „Money“ oder „Parks“) der hohe Anteil an offenen Items bemängelt. Bei Aufgaben mit einem höheren Anteil an geschlossenen Items wird häufig angegeben, dass die Distraktoren einiger Multiple-Choice-Items leicht als falsch zu erkennen seien.
- Viele Lehrkräfte äußern den Wunsch, dass ihnen die Hörtexte zukünftig auch als Transkripte zur Verfügung gestellt werden.
- In den Kommentaren der Lehrkräfte zu den Aufgaben des Kompetenzbereichs „Sprachmittlung“ wird öfter angemerkt, dass der Mediationstext zu lang sei.
- Ein hoher Anteil lobender Kommentare (z. B. im Hinblick auf das Anspruchsniveau der Aufgaben) findet sich zu den Aufgaben des Kompetenzbereichs „Schreiben“.

3.3 Französisch

Im Fach Französisch legen bundesweit betrachtet deutlich weniger Schülerinnen und Schüler eine schriftliche Abiturprüfung ab als in den Fächern Deutsch und Englisch. Die im Folgenden dargestellten Ergebnisse basieren dementsprechend auf erheblich kleineren Stichproben als die Befunde zu den anderen Fächern (vgl. Tab. 1) 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 drei Fällen wurden die Aufgaben aus dem Pool deutlich häufiger gewählt als die landeseigenen Aufgaben („Petit pays“, „Samba pour la France“ und „Voyage“ aus dem Kompetenzbereich Schreiben), in anderen Fällen zeigen sich nur geringfügige Unterschiede oder höhere Attraktivitätswerte 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):

- In der Gesamtschau unterscheiden sich die Ergebnisse, die von den Schülerinnen und Schülern insgesamt bei den Aufgaben aus dem Pool erzielt wurden, nicht signifikant von den Ergebnissen in den landeseigenen Aufgaben (g = -.17, nicht statistisch signifikant).
- Im Fach Französisch war im Rahmen der Evaluation für das Prüfungsjahr 2017 ermittelt worden, dass die Prüflinge bei den Aufgaben zum „Hörverstehen“ signifikant weniger gut abgeschnitten hatten als bei den landeseigenen Aufgaben. Für das Prüfungsjahr 2019 findet sich ein solcher Effekt nicht (g = .03, nicht statistisch signifikant).
- In den anderen beiden Kompetenzbereichen („Schreiben“ und „Sprachmittlung“) findet sich zwar deskriptiv die Tendenz, dass die Poolaufgaben eine höhere empirische Schwierigkeit aufweisen als die landeseigenen Aufgaben. Die betreffenden Unterschiede sind aber insgesamt gering und statistisch nicht signifikant (jeweils g = -.31, nicht statistisch signifikant).

Tabelle 14: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Französisch

KB[18]	Anzahl Länder	Mittlere stand. Differenz	Vertrauensbereich (95 %)
HV	8	0.03	[-0.36; 0.42]
SM	4	-0.31	[-0.76; 0.14]
S	8	-0.31	[-1.23; 0.62]
insgesamt	**20**	**-0.17**	**[-0.52; 0.17]**

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

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 = .85) und unterscheidet sich nicht statistisch signifikant von dem für die landeseigenen Aufgaben ermittelten Koeffizienten (r = .91).
- Wie im Fach Englisch wurde auch im Fach Französisch der niedrigste Trennschärfekoeffizient für die Aufgaben zum Kompetenzbereich „Hörverstehen“ gefunden (r = .72). Dieser Koeffizient ist signifikant kleiner als der für die Aufgaben zum Kompetenzbereich „Sprachmittlung“ berechnete Kennwert (r = .89), unterscheidet sich jedoch nicht signifikant von dem (allerdings auf einer sehr kleinen Aufgabenstichprobe basierenden) Trennschärfekoeffizienten, der für die Aufgaben zum Kompetenzbereich „Schreiben“ ermittelt wurde (r = .91). Auch dieses Befundmuster ähnelt insgesamt den Ergebnissen, die im Rahmen der Evaluation für das Prüfungsjahr 2017 ermittelt wurden.

Tabelle 15: Aufgabenübergreifende Ergebnisse zur Trennschärfe im Fach Französisch

KB[12]	Anzahl Länder	Trennschärfekoeffizient r	Vertrauensbereich (95 %)
Poolaufgaben			
HV	9	0.89	[0.85; 0.92]
SM	6	0.91	[0.72; 0.97]
S	9	0.72	[0.64; 0.78]
insgesamt	**24**	**0.85**	**[0.79; 0.89]**
landeseigene Aufgaben			
insgesamt	**26**	**0.91**	**[0.84; 0.95]**

3.3.4 Kriteriale 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 = .70 ermittelt. Dieser Koeffizient unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde (r = .74). Beide Koeffizienten sind als hoch einzustufen. Bei einer nach Kompetenzbereichen differenzierten Betrachtung finden sich zwar deutliche Unterschiede, diese fallen aber etwas geringer aus als bei der Evaluation für das Prüfungsjahr 2017. Für den Kompetenzbereich „Hörverstehen“ wurde ein Validitätskoeffizient mittlerer Höhe (r = .60) gefunden. Für den Kompetenzbereich „Schreiben“ wurde ein als hoch einzustufender Validitätskoeffizient (r = .69) berechnet. Beide Koeffizienten fallen signifikant geringer aus als der für den Kompetenzbereich „Sprachmittlung“ berechnete Wert (r = .77).
- Ein ähnliches Bild zeigt sich in Bezug auf das Validitätskriterium „Klausurnoten“: Der hier für die Aufgaben aus dem Pool berechnete Validitätskoeffizient (r = .71) ist ebenfalls als hoch einzustufen und unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben ermittelt wurde (r = .77). Wiederum wurde für die Aufgaben zum Kompetenzbereich „Hörverstehen“ (r = .58, mittel) ein signifikant kleinerer Validitätskoeffizient

gefunden als für die Aufgaben des Kompetenzbereichs „Sprachmittlung“ (r = .77). Für den Kompetenzbereich „Schreiben“ wurde ein von diesen beiden Werten nicht signifikant abweichender Validitätskoeffizient von r = .74 ermittelt.

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

KB[19]	Anzahl Länder	Validitätskoeffizient	Vertrauensbereich (95 %)
Poolaufgaben			
HV	9	0.77	[0.70; 0.83]
SM	6	0.69	[0.49; 0.83]
S	9	0.60	[0.48; 0.70]
insgesamt	**24**	**0.70**	**[0.63; 0.75]**
landeseigene Aufgaben			
insgesamt	**26**	**0.74**	**[0.66; 0.81]**

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	9	0.79	[0.71; 0.85]
SM	6	0.74	[0.55; 0.85]
S	9	0.58	[0.46; 0.68]
insgesamt	**24**	**0.71**	**[0.64; 0.77]**
landeseigene Aufgaben			
insgesamt	**26**	**0.77**	**[0.69; 0.83]**

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):

- Bei den Einschätzungen zum Anspruch finden sich insgesamt nur geringe Unterschiede zwischen den Aufgaben des Pools und den landeseigenen Aufgaben (g = .09, nicht statistisch signifikant).
- Auch bei einer nach Kompetenzbereichen differenzierten Betrachtung lassen sich keine statistisch signifikanten Effekte feststellen. Deskriptiv zeigt sich allerdings die Tendenz, dass die befragten Lehrkräfte die Poolaufgaben zum „Hörverstehen“ als etwas anspruchsvoller bewerten als die landeseigenen Aufgaben (g = .41, nicht statistisch signifikant). Auch im Rahmen der Evaluation für das Prüfungsjahr 2017 waren die Aufgaben zum „Hörverstehen“

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

als sehr anspruchsvoll beurteilt worden. Anders als im Prüfungsjahr 2019 war der betreffende Effekt dabei aber statistisch signifikant und als hoch einzustufen.

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

KB[20]	Anzahl Länder	Mittlere stand. Differenz	Vertrauensbereich (95 %)
HV	5	0.41	[-0.16; 0.98]
SM	2	-0.36	[-3.05; 2.33]
S	4	-0.07	[-0.98; 0.85]
insgesamt	**11**	**0.09**	**[-0.29; 0.46]**

3.3.6 Einschätzungen der Lehrkräfte zum Nutzen der Erwartungshorizonte der Aufgaben

Wie im Fach Englisch (und ähnlich der Evaluation für das Prüfungsjahr 2017) wurden auch im Fach Französisch nur geringe, statistisch nicht signifikante Unterschiede in den Einschätzungen der Lehrkräfte zu den Erwartungshorizonten der Aufgaben aus dem Pool und der landeseigenen Aufgaben gefunden (g = .03, nicht statistisch signifikant).

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

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.