Institut zur Qualitätsentwicklung
im Bildungswesen

Gemeinsame Abituraufgabenpools der Länder

Evaluation von Aufgaben der Pools für das Prüfungsjahr 2021

Ergebnisse zur Bewährung der Aufgaben

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

Inhalt

Kurzzusammenfassung

Gegenwärtig haben die Länder in vier Fächern (Deutsch, Englisch, Französisch und Mathematik) die Möglichkeit, Aufgaben aus dem gemeinsamen Abituraufgabenpool zu entnehmen und in ihren schriftlichen Abiturprüfungen einzusetzen. Der vorliegende Bericht stellt die Ergebnisse der Evaluation der Bewährung dieser Poolaufgaben in den Abiturprüfungen der Länder für das Prüfungsjahr 2021 dar, die turnusgemäß für das Fach Mathematik durchgeführt wurde.

Wie bereits im Prüfungsjahr 2020 konnten aufgrund der Pandemie die Abiturprüfungen in einigen Ländern nicht zu den vereinbarten gemeinsamen Prüfungsterminen durchgeführt werden. An der Evaluation zur Bewährung der Aufgaben aus den Pools für das Prüfungsjahr 2021 nahmen insgesamt 9 Länder teil, die sich zum Teil erheblich darin unterschieden, in welchem Umfang sie (Teil-)Aufgaben aus dem Pool in ihren Abiturprüfungen einsetzten.

Im Hinblick auf die empirische Schwierigkeit und den statistischen Zusammenhang zwischen den Prüfungsleistungen und den in der Qualifikationsphase erzielten Leistungen (im Folgenden als „kriteriale Validität“ bezeichnet) zeigen sich zwischen den Aufgaben aus den Pools und den landeseigenen Aufgaben für das Fach Mathematik kaum signifikante Unterschiede. Die für die einzelnen Länder ermittelten Befundmuster weisen jedoch eine ausgeprägte Länderspezifik auf. So lassen sich in einigen Ländern zwischen den Lösungsquoten der Poolaufgaben und den Lösungsquoten der landeseigenen Aufgaben statistisch signifikante Unterschiede feststellen, die jedoch mal zugunsten der Poolaufgaben und mal zugunsten der landeseigenen Aufgaben ausfallen. Im Hinblick auf das Evaluationskriterium der kriterialen Validität fällt auf, dass die ermittelten Zusammenhänge zwischen Prüfungsleistungen und Vorleistungen deutlich höher ausfallen als die Korrelationen, die bei der Evaluation im Prüfungsjahr 2017 festgestellt wurden. Dies könnte auf eine normierende Wirkung des Einsatzes der Poolaufgaben auf den Unterricht in der Qualifikationsphase hindeuten.

Die befragten Lehrkräfte schätzten die Poolaufgaben und die landeseigenen Aufgaben im Hinblick auf die im Rahmen der Evaluation berücksichtigten Aspekte über alle Länder hinweg gemittelt sehr ähnlich ein. Ein Blick auf die länderspezifischen Ergebnisse zeigt allerdings, dass deutliche Länderunterschiede bei der Beurteilung des Anspruchs und der Angemessenheit des Umfangs der Aufgaben bestehen.

1 Verfahrensbeschreibung

Die Evaluation der Bewährung der Aufgaben aus dem Pool für das Fach Mathematik im Prüfungsjahr 2021 basiert auf dem weiterentwickelten „Konzept zur Evaluation des Einsatzes von Aufgaben der Abituraufgabenpools für die Prüfungsjahre 2018 und 2019“, das von der Amtschefskommission „Qualitätssicherung in Schulen“ im Rahmen ihrer 89. Sitzung zustimmend zur Kenntnis genommen wurde.

Betrachtete Evaluationsbereiche

Im Bericht werden drei Evaluationsbereiche betrachtet, die in dem o. g. Evaluationskonzept spezifiziert wurden und in den jeweiligen Abschnitten dieses Dokuments erläutert werden:

- Empirische Schwierigkeit der Aufgaben (Unterscheidet sich die empirische Schwierigkeit der Aufgaben aus dem Pool von der empirischen Schwierigkeit der landeseigenen Aufgaben?)
- Kriteriale Validität der Aufgaben (Gibt es einen Zusammenhang zwischen den Leistungen in den Abiturprüfungen und den Vorleistungen[1] der Prüflinge?)[2]
- Lehrkräfteeinschätzungen zu den Aufgaben (Wie werden die einzelnen Prüfungsaufgaben, die entweder aus dem Pool für das Fach Mathematik stammen oder von den Ländern selbst entwickelt wurden, von Lehrkräften eingeschätzt?) in Bezug auf folgende Aspekte:
  - Anspruch der Aufgaben
  - Nützlichkeit der Erwartungshorizonte
  - Angemessenheit des Umfangs der Aufgaben
  - Sprachliche Eindeutigkeit und Verständlichkeit der Aufgabenstellungen
  - Verständlichkeit der Aufgabentexte

Stichprobe und Datenerhebung

An der Evaluation zur Bewährung der Aufgaben aus dem Pool für das Prüfungsjahr 2021 nahmen 9 Länder teil. Das für die Datenerhebung vom IQB bereitgestellte elektronische Eingabeinstrument wurde von 8 Ländern genutzt. Das Land Nordrhein-Westfalen übermittelte Daten aus der landeseigenen Erhebung. In den Ländern Bremen und Hamburg 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 Nordrhein-Westfalen, Niedersachsen und dem Saarland wurden die Schulen, an denen die für die Evaluation benötigten Daten erhoben wurden, vom jeweils zuständigen Landesinstitut ausgewählt. Im Land Bremen wurde durch das zuständige Landesinstitut eine flächendeckende Beteiligung aller Gymnasien und aller allgemeinbildenden Schulen mit einer

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

[2] Darüber hinaus wurde analysiert, wie stark die bei einzelnen Aufgaben erzielten Ergebnisse mit dem Gesamtergebnis der Prüfung zusammenhängen. In Bezug auf dieses, im Evaluationskonzept als Trennschärfe bezeichnete Kriterium, wurden detaillierte, länderscharfe Analysen auf Teilaufgabenebene durchgeführt. Die Befunde dieser Detailanalysen werden, analog zum Vorgehen bei der im Prüfungsjahr 2017 durchgeführten Evaluation, den Mitgliedern der AG Aufgaben Mathematik zur Verfügung gestellt. Auf eine zusammenfassende, aggregierte Darstellung der Ergebnisse zur Trennschärfe der Aufgaben wird im Rahmen dieses Berichts verzichtet, da sie inhaltlich nur schwer zu interpretieren wäre.

gymnasialen Oberstufe initiiert. In den anderen Ländern wurde die Evaluationsstichprobe in Abstimmung mit den jeweiligen Ansprechpartnerinnen und -partnern der Länder durch das IQB gezogen. Dazu wurden die Länder gebeten, dem IQB anonymisierte Listen aller infrage kommenden Schulen zur Verfügung zu stellen. Auf dieser Grundlage erfolgte eine zufällige Ziehung der Schulstichproben, die pro Land 20 Schulen umfassen. 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 Anforderungsniveau ein Kurs in die Datenerhebung einbezogen, sofern in der entsprechenden Abiturprüfung Aufgaben aus dem Pool für das Fach Mathematik eingesetzt wurden und Schülerinnen und Schüler eine Prüfung abgelegt haben. Damit waren maximal 2 Kurse pro Schule an der Datenerhebung beteiligt. 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.[3,4] Insgesamt wurden in die Evaluation Daten zu 7047 Prüfungsarbeiten, davon 3904 Prüfungsarbeiten auf erhöhtem und 3143 Prüfungsarbeiten auf grundlegendem Anforderungsniveau, einbezogen.

Datenauswertung

Die Ergebnisse der Evaluation der Bewährung der Poolaufgaben für das Fach Mathematik werden gegliedert nach Ländern (in anonymisierter Form) dargestellt. Dabei werden – für jedes Sachgebiet und jedes zugelassene digitale Hilfsmittel – jeweils getrennt nach Anforderungsniveaus die Teilaufgaben aus dem Pool mit den landeseigenen Teilaufgaben verglichen. Die Darstellung der Ergebnisse erfolgt anonymisiert, indem die Ergebnisse jeweils für „Land 1“, „Land 2“, „Land 3“ usw. ausgewiesen werden. Die Reihenfolge der Länder wurde zufällig gewählt, ist aber identisch mit der Reihenfolge in den anderen Teilen des vorliegenden Evaluationsberichts.

In den Evaluationsbereichen „Auswahl der Aufgaben“ und „Schwierigkeit der Aufgaben“ sowie für die Lehrkräfteeinschätzungen wurden zunächst deskriptive Kennwerte (d. h. relative Häufigkeiten und Mittelwerte) berechnet, inferenzstatistisch auf signifikante Unterschiede geprüft[5] und in Form von Grafiken dargestellt. Im Bereich „Validität der Aufgaben“ wurde die Korrelation der bei einer Aufgabe erzielten Ergebnisse sowohl bezüglich der in der Qualifikationsphase erzielten Klausurnoten als auch bezüglich der Halbjahresnoten bestimmt.

Analog zum Vorgehen zur Evaluation für das Prüfungsjahr 2019 wurden die deskriptiven Detailergebnisse zur Bewährung der einzelnen Aufgaben außerdem mittels metaanalytischer Methoden zu länderübergreifenden Gesamtergebnissen zusammengefasst. Hierbei wird jedes Einzelergebnis (d. h. jeder Kennwert, der für eine Aufgabe aus dem Pool in einem einzelnen Land berechnet wurde) als eine „Studie“ in die Auswertung einbezogen. Zudem wurden

[3] 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)“.

[4] Diese Beschränkung wurde für die Länder Mecklenburg-Vorpommern und Bremen nicht angewandt.

[5] Die Signifikanzprüfung erfolgte mithilfe sogenannter Mixed-Effects-Modelle, um die aus der Stichprobenziehung resultierende Datenstruktur zu berücksichtigen. Der Grund dafür liegt darin, dass die erhobenen Daten in Clustern geschachtelt sind, bei denen es sich um die verschiedenen Kurse handelt. Da der Stichprobenfehler bei einer solchen Schachtelung höher ist als bei einer einfachen Zufallsauswahl, müssen bei der statistischen Datenauswertung Methoden verwendet werden, die dieser Datenstruktur Rechnung tragen.

Differenzierungen nach Anforderungsniveau und Sachgebiet vorgenommen.[6] Als Ergebnis der Metaanalysen wird jeweils ein über alle Länder, Anforderungsniveaus und Poolaufgaben zusammengefasster Effekt berechnet, der zum Beispiel angibt, ob die Aufgaben aus dem Pool (nach Einschätzung der Lehrkräfte) in einem bestimmten Fach insgesamt eher weniger anspruchsvoll oder eher anspruchsvoller als die landeseigenen Aufgaben waren.

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

2 Empirische Schwierigkeit der Aufgaben im Fach Mathematik

Im Fach Mathematik wurde die empirische Aufgabenschwierigkeit anhand der Lösungsquote bestimmt, also anhand des Quotienten aus der durchschnittlichen Anzahl der von den Prüflingen in den betreffenden Teilaufgaben erreichten Bewertungseinheiten (BE) und der Anzahl der maximal erreichbaren BE. Die zu jedem Land ermittelten Schwierigkeitskennwerte wurden mittels metaanalytischer Methoden zu standardisierten Effektmaßen (*Hedge's g*) aggregiert. In den Tabellen 1 und 2 sind die Ergebnisse dieser Metaanalysen dargestellt. Hierfür wurden aufgabenübergreifend für jedes Land die Differenzen zwischen der mittleren empirischen Schwierigkeit der Teilaufgaben aus dem Pool und der mittleren empirischen Schwierigkeit der landeseigenen Teilaufgaben ermittelt. Die Differenzen wurden standardisiert und über alle Länder aggregiert, die an der Evaluation teilgenommen haben. In einer weiteren Analyse wurden Unterschiede in den Lösungsquoten bei Teilaufgaben aus dem Pool und landeseigenen Teilaufgaben getrennt nach Sachgebieten betrachtet, wobei die ohne Hilfsmittel zu bearbeitenden Prüfungsteile separat ausgewertet wurden. Als standardisiertes Effektmaß für diese Unterschiede 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). *Hedges' g* sowie die untere und obere Grenze des Vertrauensbereichs[7] sind in den folgenden Tabellen aufgabenübergreifend sowie separat für jedes Sachgebiet dargestellt. Eine negative mittlere Differenz bedeutet, dass Prüflinge bei der Bearbeitung von Teilaufgaben aus dem Pool im Durchschnitt weniger gute Ergebnisse erzielten als bei der Bearbeitung landeseigener Teilaufgaben. Eine positive Differenz weist darauf hin, dass Prüflinge bei Teilaufgaben aus dem Pool bessere Ergebnisse erzielten. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 1: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Mathematik

Sachgebiet	Anzahl der „Studien"	Mittlere stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
insgesamt	26	-0.07	[-0.17; 0.04]

Tabelle 2: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Sachgebieten im Fach Mathematik

Sachgebiete[8]	Anzahl der „Studien"	Mittlere stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
HF	22	0.08	[-0.28; 0.11]
A	14	0.22	[-0.06; 0.50]
AG	11	-0.23*	[-0.46; -0.01]
S	7	0.32	[-0.04; 0.69]
LA	2	1.38	[-2.39; 5.16]

[7] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.

[8] HF = Hilfsmittelfreier Teil, A = Analysis, AG = Analytische Geometrie, S = Stochastik, LA = Lineare Algebra.

Die Ergebnisse der Metanalysen zur empirischen Aufgabenschwierigkeit lassen sich wie folgt zusammenfassen:

- Insgesamt unterscheiden sich die Lösungsquoten für die Teilaufgaben aus dem Pool statistisch nicht signifikant von den landeseigenen Aufgaben (g = -.07). Ein Blick auf die länderspezifischen Ergebnisse zeigt allerdings, dass die in Bezug auf die empirische Aufgabenschwierigkeit ermittelten Befundmuster eine ausgeprägte Länderspezifik aufweisen. So sind in einigen Ländern statistisch signifikante Unterschiede zwischen den Lösungsquoten festzustellen, die allerdings mal zugunsten der Poolaufgaben und mal zugunsten der landeseigenen Aufgaben ausfallen.
- In einer nach Sachgebieten differenzierten Analyse zeigt sich im Prüfungsjahr 2021 ein statistisch signifikanter Unterschied im Sachgebiet „Analytische Geometrie“. Hier wurden bei den Teilaufgaben aus dem Pool signifikant geringere Lösungsquoten erzielt als bei den landeseigenen Aufgaben. Die Höhe des gefundenen Effekts ist allerdings als „klein“ einzustufen (g = -0.23). Anders als bei der Evaluation im Prüfungsjahr 2017 zeigt sich im Prüfungsjahr 2021 für das Sachgebiet „Stochastik“ kein statistisch signifikanter Unterschied zwischen den Lösungsquoten der Teilaufgaben aus dem Pool und den Lösungsquoten der landeseigenen Aufgaben (g = 0.32). Auch für die übrigen Sachgebiete wurden keine statistisch bedeutsamen Unterschiede ermittelt.

3 Kriteriale Validität der Aufgaben im Fach Mathematik

Als Indikatoren für die kriteriale Validität[9] einer Aufgabe wurden die Korrelationen der bei der Aufgabe erzielten Ergebnisse mit den in der Qualifikationsphase erreichten Klausurnoten und Halbjahresnoten bestimmt. Analog zum Vorgehen bei der Metaanalyse zur empirischen Schwierigkeit wurden die ermittelten Korrelationskoeffizienten für jedes Sachgebiet über jene Länder aggregiert, die an der Evaluation teilgenommen haben. Die so berechneten Validitätskoeffizienten sowie die untere und obere Grenze des Vertrauensbereichs[10] sind in den Tabellen 3 und 4 separat für jedes Sachgebiet und über alle Sachgebiete hinweg dargestellt. Die Höhe der 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 $r = 0{,}60$ als „mittel“ und Koeffizienten ab $r = 0{,}60$ als „hoch“ (vgl. Fisseni, 1997). Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 3: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Mathematik

Sachgebiete[11]	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben			
HF	26	0.68	[0.63; 0.73]
A	20	0.69	[0.66; 0.73]
AG	19	0.69*	[0.65; 0.73]
S	8	0.56	[0.49; 0.63]
LA	3	0.74	[0.70; 0.77]
insgesamt	**26**	**0.77**	**[0.74; 0.80]**
landeseigene Aufgaben			
HF	21	0.67	[0.62; 0.72]
A	15	0.73	[0.66; 0.78]
AG	11	0.64*	[0.58; 0.69]
S	11	0.66	[0.55; 0.75]
LA	1	0.69	[0.15; 0.91]
insgesamt	**26**	**0.74**	**[0.69; 0.78]**

[9] Die kriteriale Validität der Aufgaben ist ein Maß für den Zusammenhang zwischen den Leistungen in der Abiturprüfung und den Vorleistungen der Prüflinge. Der Zusammenhang ist umso größer je größer der Validitätskoeffizient ist.

[10] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.

[11] HF = Hilfsmittelfreier Teil, A = Analysis, AG = Analytische Geometrie, S = Stochastik, LA = Lineare Algebra.

Tabelle 4: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Mathematik

Sachgebiete[12]	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben			
HF	20	0.68	[0.63; 0.72]
A	14	0.71	[0.66; 0.75]
AG	15	0.70	[0.67; 0.74]
S	7	0.58	[0.52; 0.63]
LA	1	0.81	[0.51; 0.93]
insgesamt	**20**	**0.78**	**[0.74; 0.81]**
landeseigene Aufgaben			
HF	15	0.66	[0.62; 0.70]
A	9	0.68	[0.63; 0.73]
AG	7	0.66	[0.62; 0.70]
S	11	0.67	[0.52; 0.78]
LA	-	-	
insgesamt	**20**	**0.72**	**[0.67; 0.77]**

Die als Indikatoren zur kriterialen Validität der Aufgaben im Fach Mathematik ermittelten Ergebnisse lassen sich wie folgt zusammenfassen (vgl. Anhang 2, Abschnitt 4):

- Der in Bezug auf das Validitätskriterium „Halbjahresnoten“ für die Teilaufgaben aus dem Pool berechnete Validitätskoeffizient ist als hoch einzustufen (r = .77) und unterscheidet sich nicht statistisch signifikant von dem Kennwert, der für die landeseigenen Teilaufgaben gefunden wurde (r = .74). Der kleinste Validitätskoeffizient wurde für die Teilaufgaben aus dem Pool zum Sachgebiet „Stochastik“ ermittelt (r = .56, als mittel einzustufen). Ein statistisch signifikanter (aber dennoch geringer) Unterschied zwischen den Poolaufgaben und den landeeigenen Aufgaben wurde nur im Sachgebiet „Analytische Geometrie“ festgestellt (Poolaufgaben, r = .69, landeseigene Aufgaben: r = .64, jeweils als hoch einzustufen).
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Teilaufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = .78 (als hoch einzustufen) ermittelt. Dieser Koeffizient unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde (r = .72, als hoch einzustufen). Der kleinste Validitätskoeffizient wurde wiederum für die Teilaufgaben aus dem Pool zum Sachgebiet „Stochastik“ ermittelt (r = .58, als mittel einzustufen). Im Vergleich der für die einzelnen Sachgebiete berechneten Validitätskoeffizienten zeigen sich keine statistisch signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben.

[12] HF = Hilfsmittelfreier Teil, A = Analysis, AG = Analytische Geometrie, S = Stochastik, LA = Lineare Algebra.

Hervorzuheben ist, dass die für das Prüfungsjahr 2021 ermittelten Validitätskoeffizienten deutlich höher als im Prüfungsjahr 2017 ausfallen. Dies könnte darauf hindeuten, dass der Einsatz der Poolaufgaben in den Abiturprüfungen der Länder eine positive, normierende Rückwirkung auf den Unterricht und die Aufgabengestaltung in der Qualifikationsphase hat.[13]

4 Lehrkräfteeinschätzungen zu den Aufgaben im Fach Mathematik

Im Rahmen der Evaluation zur Bewährung der Poolaufgaben wurden die Lehrkräfte der für die Evaluation ausgewählten Kurse gebeten, die Abiturprüfungsaufgaben im Hinblick auf die folgenden fünf Aspekte einzuschätzen:

- Anspruch der Aufgaben
- Nützlichkeit der Erwartungshorizonte
- Angemessenheit des Umfangs der Aufgaben
- Sprachliche Eindeutigkeit und Verständlichkeit der Aufgabenstellungen
- Verständlichkeit der Aufgabentexte

Die Einschätzungen zu den drei letztgenannten Aspekten wurden erfasst, um auf diesem Wege erste Informationen darüber zu erhalten, wie Lehrkräfte die Bewältigbarkeit von Prüfungsaufgaben im Fach Mathematik beurteilen.

Die Lehrkräfteeinschätzungen zu den fünf genannten Aspekten wurden nicht spezifisch für Teilaufgaben, sondern nur bezogen auf vollständige Aufgaben der Abiturprüfung erhoben, die im Falle von Aufgaben, die auf Poolaufgaben basieren, nicht ausschließlich aus Teilaufgaben aus dem Pool bestanden und teilweise modifizierte Teilaufgaben aus dem Pool enthielten. . Eine Interpretation der Ergebnisse in Bezug auf die aus dem Pool eingesetzten Teilaufgaben ist somit nur eingeschränkt möglich. Eine Aufgabe bzw. ein Prüfungsteil wurde dabei immer dann als aus dem Pool stammend betrachtet, wenn die darin enthaltenen Teilaufgaben im Umfang von mehr als der Hälfte der maximal erreichbaren Bewertungseinheiten aus dem Pool entnommen wurden.

[13] In der Forschungsliteratur werden solche Rückwirkungen auch als „washback“- bzw. „backwash“-Effekte bezeichnet.

4.1 Anspruch der Aufgaben

In der Tabelle 5 sind die Ergebnisse der Metaanalyse für die Einschätzungen der Lehrkräfte zum Anspruch der Aufgaben dargestellt. Differenzen zwischen den Einschätzungen zu den Aufgaben aus dem Pool einerseits und zu den landeseigenen Aufgaben andererseits wurden anhand des standardisierten Effektmaßes *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). Ein negatives Vorzeichen zeigt an, dass der Anspruch der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als geringer beurteilt wurde. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte den Anspruch der Poolaufgaben höher einschätzten als den Anspruch der landeseigenen Aufgaben. Für das berechnete Effektmaß sind zusätzlich die untere und obere Grenze des Vertrauensbereichs[14] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 5: Aufgabenübergreifende Ergebnisse zum eingeschätzten Anspruch der Aufgaben im Fach Mathematik

Sachgebiete	Anzahl der „Studien"	Mittlere stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
insgesamt	17	0.21	[-0.06; 0.48]

Im Vergleich zu den landeseigenen Aufgaben haben die Lehrkräfte den Anspruch der Poolaufgaben im Fach Mathematik insgesamt also etwas höher eingeschätzt (g = .21). Der Unterschied ist jedoch sehr gering und statistisch nicht signifikant.

[14] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.

4.2 Erwartungshorizont der Aufgaben

In der Tabelle 6 sind die Ergebnisse der Metaanalyse für die Einschätzungen der Lehrkräfte zu den Erwartungshorizonten der Aufgaben dargestellt. Differenzen zwischen den Einschätzungen zu den Aufgaben aus dem Pool einerseits und zu den landeseigenen Aufgaben andererseits wurden anhand des standardisierten Effektmaßes *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). Ein negatives Vorzeichen zeigt an, dass der Nutzen der Erwartungshorizonte der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als geringer beurteilt wurde. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte den Nutzen der Erwartungshorizonte der Poolaufgaben höher einschätzten als den Nutzen der landeseigenen Aufgaben. Für das berechnete Effektmaß sind zusätzlich die untere und obere Grenze des Vertrauensbereichs[15] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 6: Einschätzungen der Lehrkräfte zu den Erwartungshorizonten der Aufgaben im Fach Mathematik

Sachgebiete	Anzahl der „Studien"	Mittlere stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
insgesamt	17	-0.03	[-0.15; 0.09]

Insgesamt wurden die Erwartungshorizonte der Poolaufgaben und der landeseigenen Aufgaben also nahezu identisch beurteilt ($g = -.03$).

[15] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.

4.3 Umfang der Aufgaben

In der Tabelle 7 sind die Ergebnisse der Metaanalyse für die Einschätzungen der Lehrkräfte zur Angemessenheit des Umfangs der Aufgaben dargestellt. Differenzen zwischen den Einschätzungen zu den Aufgaben aus dem Pool einerseits und zu den landeseigenen Aufgaben andererseits wurden anhand des standardisierten Effektmaßes *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). Ein negatives Vorzeichen zeigt an, dass der Umfang der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als weniger angemessen beurteilt wurde. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte den Umfang der Poolaufgaben als gemessener einschätzten als den Umfang der landeseigenen Aufgaben. Für das berechnete Effektmaß ist zusätzlich die untere und obere Grenze des Vertrauensbereichs[16] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 7: Einschätzungen der Lehrkräfte zur Angemessenheit des Umfangs der Aufgaben im Fach Mathematik

Sachgebiete	Anzahl der „Studien"	Mittlere stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
insgesamt	17	-0.17	[-0.37; 0.03]

Im Vergleich zu den landeseigenen Aufgaben haben die Lehrkräfte den Umfang der Poolaufgaben im Fach Mathematik insgesamt etwas weniger positiv beurteilt (g = -.17). Dieser Unterschied ist jedoch sehr gering und statistisch nicht signifikant.

[16] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.

4.4 Sprachliche Eindeutigkeit und Verständlichkeit der Aufgabenstellungen

In der Tabelle 8 sind die Ergebnisse der Metaanalyse für die Einschätzungen der Lehrkräfte zur sprachlichen Eindeutigkeit und Verständlichkeit der Arbeitsaufträge der Aufgaben dargestellt. Differenzen zwischen den Einschätzungen zu den Aufgaben aus dem Pool einerseits und zu den landeseigenen Aufgaben andererseits wurden anhand des standardisierten Effektmaßes *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). Ein negatives Vorzeichen zeigt an, dass die Arbeitsaufträge der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als sprachlich weniger klar und verständlich beurteilt wurden. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte die Arbeitsaufträge der Poolaufgaben als sprachlich klarer und verständlicher einschätzten als die Arbeitsaufträge der landeseigenen Aufgaben. Für das berechnete Effektmaß ist zusätzlich die untere und obere Grenze des Vertrauensbereichs[17] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 8: Einschätzungen der Lehrkräfte zur sprachlichen Eindeutigkeit und Verständlichkeit der Aufgaben im Fach Mathematik

Sachgebiete	Anzahl der „Studien"	Mittlere stand. Differenz *(Hedges' g)*	Vertrauensbereich (95%)
insgesamt	17	0.01	[-0.15; 0.17]

Insgesamt wurden die Arbeitsaufträge der Poolaufgaben und der landeseigenen Aufgaben also nahezu identisch beurteilt ($g = .01$).

[17] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.

4.5 Verständlichkeit der Aufgabentexte

In der Tabelle 9 sind die Ergebnisse der Metaanalyse für die Einschätzungen der Lehrkräfte zur Verständlichkeit der Aufgabentexte dargestellt. Differenzen zwischen den Einschätzungen zu den Aufgaben aus dem Pool einerseits und zu den landeseigenen Aufgaben andererseits wurden anhand des standardisierten Effektmaßes *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). Ein negatives Vorzeichen zeigt an, dass die Aufgabentexte der Poolaufgaben im Vergleich zu den Aufgabentexten der landeseigenen Aufgaben als weniger verständlich beurteilt wurden. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte die Aufgabentexte der Poolaufgaben als verständlicher einschätzten als die Aufgabentexte der landeseigenen Aufgaben. Für das berechnete Effektmaß ist zusätzlich die untere und obere Grenze des Vertrauensbereichs[18] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 9: Einschätzungen der Lehrkräfte zur Verständlichkeit der Aufgabentexte

Sachgebiete	Anzahl der „Studien"	Mittlere stand. Differenz *(Hedges' g)*	Vertrauensbereich (95%)
insgesamt	17	-0.03	[-0.23; 0.16]

Insgesamt wurden die Aufgabentexte der Poolaufgaben und der landeseigenen Aufgaben nahezu identisch beurteilt ($g = -.03$).

[18] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.