Gemeinsame Abituraufgabenpools der Länder

Evaluation des Einsatzes von Aufgaben der Pools für das Prüfungsjahr 2023

Ergebnisse zur Bewährung der Aufgaben

Dr. Lars Hoffmann, Dr. Lea Kröger, Dr. Marco Schickel, 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 zum Einsatz von Aufgaben der gemeinsamen Abituraufgabenpools der Länder für das Prüfungsjahr 2023 dar. Hierbei werden folgende Bereiche betrachtet:

- eine Befragung der Lehrkräfte zu den eingesetzten Poolaufgaben
- eine Erhebung von Ergebnissen zu den Vor- und Prüfungsleistungen der Schülerinnen und Schüler im Fach Mathematik

Befragung der Lehrkräfte

In den Fächern Deutsch, Englisch und Französisch fallen die Rückmeldungen der befragten Lehrkräfte zu den Poolaufgaben in Bezug auf die abgefragten Aspekte (z. B. Schwierigkeitsgrad der Aufgaben, Klarheit und Verständlichkeit der Aufgabenstellungen) insgesamt positiv aus. Für das Fach Mathematik ergeben die Ergebnisse der Lehrkräftebefragung ein heterogenes Bild: Während die Aufgaben des Prüfungsteils A im Hinblick auf die abgefragten Aspekte (z. B. Dichte und Schwierigkeitsgrad der Aufgaben, Angemessenheit der sprachlichen Komplexität der Aufgabenstellungen) überwiegend positiv eingeschätzt wurden, beurteilten die befragten Lehrkräfte die Aufgaben des Prüfungsteils B deutlich kritischer.

Ergebnisse zur Erhebung der Vor- und Prüfungsleistungen von Schülerinnen und Schülern im Fach Mathematik

Die ermittelten Kennwerte zur empirischen Schwierigkeit der Aufgaben zeigen, dass die Prüflinge bei den Poolaufgaben weniger gut abgeschnitten haben als bei den landeseigenen Aufgaben. Die festgestellten Unterschiede sind allerdings nicht in jedem Fall statistisch signifikant und fallen vor allem beim Prüfungsteil B geringer aus als es die Ergebnisse der Lehrkräftebefragung vermuten lassen. Im Mittel sind die statistischen Zusammenhänge zwischen den Prüfungsleistungen und den in der Qualifikationsphase erzielten Leistungen (im Folgenden als „kriteriale Validität“ bezeichnet) für die Poolaufgaben aller Sachgebiete als hoch einzustufen, wobei die Prüflinge in der Qualifikationsphase signifikant mehr Notenpunkte erzielten als in der Abiturprüfung.

1 Verfahrensbeschreibung

Die Evaluation, deren Ergebnisse im vorliegenden Bericht dokumentiert sind, basiert auf einem Evaluationskonzept, das in Abstimmung mit den jeweils zuständigen Gremien der KMK, insbesondere mit der AG Abiturkommission und der Amtschefskommission „Qualitätssicherung in Schulen“, bzw. mit dem Kuratorium des IQB kontinuierlich entwickelt wurde. Der vorliegende Bericht beinhaltet die Ergebnisse zu zwei zentralen Bestandteilen der ersten Säule des Konzepts für das Prüfungsjahr 2023:

(1) Die Befragung der Lehrkräfte wurde (wie im Konzept vorgesehen) in allen vier Fächern durchgeführt, in denen Poolaufgaben zur Verfügung stehen. Die Ergebnisse dieser Befragung finden sich in den Abschnitten 2.1 (Deutsch), 2.2 (Englisch), 2.3 (Französisch) und 2.4 (Mathematik)

(2) Die Erhebung von Daten zu den Vor- und Prüfungsleistungen von Schülerinnen und Schülern erfolgte turnusgemäß in dem Fach Mathematik, nicht jedoch in den drei sprachlichen Fächern (Deutsch, Englisch und Französisch). Auf der Grundlage dieser Daten wurden die folgenden Bereiche im vorliegenden Bericht betrachtet:

- Empirische Schwierigkeit der Aufgaben (Wie erfolgreich werden die Aufgaben aus den Pools bearbeitet? Unterscheidet sich die Schwierigkeit der Aufgaben aus den Pools von der Schwierigkeit der landeseigenen Aufgaben? Unterscheiden sich die Aufgaben aus den Pools und die landeseigenen Aufgaben auch dann hinsichtlich ihrer Schwierigkeit, wenn die Vorleistungen der Prüflinge berücksichtigt werden?)
    - Die Ergebnisse hierzu finden sich in dem Abschnitt 3.1
- Kriteriale Validität der Aufgaben in Bezug auf Vorleistungen (Gibt es einen Zusammenhang zwischen den Prüfungsleistungen und den Vorleistungen der Schülerinnen und Schüler? Unterscheiden sich Poolaufgaben und landeseigene Aufgaben hinsichtlich der kriterialen Validität?)
    - Die Ergebnisse hierzu finden sich in dem Abschnitt 3.2

Stichprobe und Datenerhebung

An der Evaluation für das Prüfungsjahr 2023 nahmen insgesamt 16 Länder teil. Das für die Datenerhebung vom IQB bereitgestellte elektronische Eingabeinstrument wurde (zumindest teilweise) von insgesamt 14 Ländern genutzt. Zwei Länder übermittelten Daten aus ihren landeseigenen Erhebungen. Nicht in allen der 16 teilnehmenden Länder erfolgten die Befragung der Lehrkräfte sowie die Erhebung von Daten zu Vor- und Prüfungsleistungen im gleichen Umfang. Im Fach Französisch wurde in den Ländern, in denen nur sehr wenige Schülerinnen und Schüler eine schriftliche Abiturprüfung ablegen, auf eine Erhebung verzichtet. In den Ländern, in denen das IQB die Datenerhebung durchführte, wurde in der Regel eine 20 Schulen umfassende Schulstichprobe für die Evaluation ausgewählt. 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. Im Fach Mathematik wurde an jeder Schule für jedes Anforderungsniveau ein Kurs in die Datenerhebung einbezogen, sofern in den entsprechenden Abiturprüfungen Aufgaben aus den Pools eingesetzt wurden und Schülerinnen und Schüler eine Prüfung abgelegt haben. Damit waren maximal zwei Kurse pro Schule an der Datenerhebung beteiligt. Die Auswahl der Kurse erfolgte jeweils durch Verantwortliche der

Schulleitung. Darüber hinaus wurden in jedem Fach pro Schule und Anforderungsniveau jeweils zwei Lehrkräfte um eine Teilnahme an der Befragung gebeten. Die nachfolgende Tabelle gibt einen Überblick über die Anzahl der Prüflinge und Lehrkräfte, die in die Evaluation einbezogen wurden.

Tabelle 1: Überblick über die der Evaluation zugrundeliegende Stichprobe

	Anzahl der Prüflinge		Anzahl der Lehrkräfte	
	Erhöhtes Niveau	Grundlegendes Niveau	Erhöhtes Niveau	Grundlegendes Niveau
Deutsch	-	-	369	201
Englisch	-	-	514	411
Französisch	-	-	169	-
Mathematik	6.155	3.455	518	274

Datenauswertung

Für den vorliegenden Bericht wurden die für jedes einzelne Land ermittelten Evaluationsergebnisse (analog zum Vorgehen bei der Evaluation für die vorherigen Prüfungsjahre) 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 Differenzierungen nach Anforderungsniveaus und Aufgabenarten bzw. Kompetenzbereichen vorgenommen.[1]. Als Ergebnis der Metaanalysen wird jeweils ein über alle Länder, Anforderungsniveaus und Poolaufgaben zusammengefasster Effekt berechnet, der zum Beispiel angibt, ob bei den Aufgaben aus dem Pool in einem bestimmten Fach insgesamt eher bessere oder eher weniger gute Ergebnisse erzielt wurden als bei den landeseigenen Aufgaben. Weitere Details zu diesen Vergleichen und zur Interpretation der Diagramme werden in den entsprechenden Abschnitten erläutert.

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

2 Befragung der Lehrkräfte

In allen vier Fächern wurden Lehrkräfte an den für die Evaluation ausgewählten Schulen gebeten, die Abiturprüfungsaufgaben unter fachspezifischen Aspekten einzuschätzen. Dabei dürften sie in der Regel nicht gewusst haben, bei welchen der Aufgaben es sich um landeseigene oder um Poolaufgaben handelte. Diese Einschätzungen erfolgten vorrangig anhand geschlossener Fragebogenitems („Multiple-Choice-Fragen"). In den folgenden Abschnitten werden die Ergebnisse der länderübergreifenden Analysen zur Lehrkräftebefragung zusammengefasst.

2.1 Deutsch

In den folgenden TabellenTabelle 2: Ergebnisse zum Fragebogenitem „Der Schwierigkeitsgrad der Aufgabe ist insgesamt…" (5-stufig, deutlich zu niedrig … deutlich zu hoch)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	285	3.25 (0.11)	3.28 (0.13)	0.04	[-0.35; 0.44]
EP	179	3.30 (0.22)	3.33 (0.14)	0.30	[-0.47; 1.08]
AP	116	3.34 (0.26)	3.19 (0.10)	0.17	[-0.84; 1.17]
EL	145	3.21 (0.17)	3.09 (0.10)	0.41	[-0.16; 0.97]
MI	32	2.88 (0.00)	3.99 (0.00	-1.52*	[-1.99; -1.05]
MA	84	3.60 (0.23)	3.18 (0.20)	0.61	[-0.34; 1.56]
Anforderungsniveau					
EN	216	3.31 (0.16)	3.28 (0.16)	0.36	[-0.03; 0.76]
GN	161	3.28 (0.25)	3.26 (0.12)	-0.01	[-0.28; 0.26]
insgesamt	**377**	**3.30 (0.20)**	**3.27 (0.14)**	**0.20**	**[-0.05; 0.46]**

Tabelle 3: Ergebnisse zum Fragebogenitem „Die Bewertungshinweise ermöglichen eine angemessene Einschätzung der Prüfungsleistung.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	280	3.11 (0.17)	3.07 (0.10)	0.07	[-0.14; 0.27]
EP	120	2.81 (0.14)	2.94 (0.11)	-0.15	[-0.51; 0.21]
AP	118	3.13 (0.19)	3.10 (0.08)	0.09	[-0.19; 0.37]
EL	85	3.09 (0.18)	3.15 (0.10)	-0.10	[-0.71; 0.50]
MI	30	3.13 (0.00)	2.66 (0.00)	0.63*	[0.19; 1.08]
MA	85	2.91 (0.08)	3.16 (0.08)	-0.38*	[-0.60; -0.15]
Anforderungsniveau					
EN	188	3.08 (0.16)	3.08 (0.09)	-0.06	[-0.24; 0.12]
GN	131	2.96 (0.15)	3.00 (0.11)	0.01	[-0.15; 0.18]
insgesamt	**319**	**3.04 (0.17)**	**3.05 (0.10)**	**-0.02**	**[-0.15; 0.10]**

Tabelle 4: Ergebnisse zum Fragebogenitem „Die Aufgabenstellung ist klar und verständlich formuliert.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	282	3.67 (0.15)	3.48 (0.14)	0.27	[-0.00; 0.55]
EP	120	3.47 (0.15)	3.47 (0.14)	-0,01	[-0.30; 0.27]
AP	116	3.65 (0.12)	3.38 (0.17)	0,44	[-0.22; 0.11]
EL	86	3.33 (0.13)	3.36 (0.17)	-0,11	[-0.71; 0.48]
MI	30	3.57 (0.00)	3.30 (0.00)	0,39	[-0.04; 0.83]
MA	84	3.48 (0.11)	3.60 (0.17)	-0,23	[-0.67; 0.21]
Anforderungsniveau					
EN	188	3.57 (0.16)	3.54 (0.11)	0.03	[-0.15; 0.21]
GN	130	3.55 (0.15)	3.39 (0.21)	0.32*	[0.02; 0.61]
insgesamt	**318**	**3.56 (0.16)**	**3.46 (0.15)**	**0.13**	**[-0.03; 0.29]**

Tabelle 5: Ergebnisse zum Fragebogenitem „Für die Bearbeitung der Aufgabe ist domänenspezifisches Wissen erforderlich.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	277	3.45 (0.11)	3.49 (0.10)	-0.05	[-0.22; 0.13]
EP	121	3.12 (0.13)	3.48 (0.15)	-0.53*	[-0.84; -0.22]
AP	111	3.41 (0.09)	3.50 (0.05)	-0.15	[-0.62; 0.33]
EL	83	3.59 (0.11)	3.38 (0.17)	0.30	[-0.06; 0.66]
MI	30	3.10 (0.00)	3.54 (0.00)	-0.63*	[-1.07; -0.18]
MA	84	3.46 (0.16)	3.29 (0.22)	-0.02	[-1.13; 1.10]
Anforderungsniveau					
EN	184	3.45 (0.14)	3.42 (0.18)	-0.07	[-0.30; 0.16]
GN	128	3.28 (0.14)	3.38 (0.11)	-0.18	[-0.40; 0.03]
insgesamt	**312**	**3.39 (0.14)**	**3.45 (0.14)**	**-0.11**	**[-0.27; 0.05]**

Tabelle 6: Ergebnisse zum Fragebogenitem „Der Text bzw. die Texte der Aufgabe sind angemessen im Hinblick auf die sprachliche Komplexität."
(4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	278	3.46 (0.12)	3.46 (0.13)	0.00	[-0.24; 0.23]
EP	120	3.11 (0.16)	3.41 (0.19)	-0.56	[-1.21; 0.08]
AP	112	3.21 (0.13)	3.46 (0.08)	-0.30	[-0.66; 0.05]
EL	83	3.39 (0.17)	3.45 (0.09)	-0.28	[-0.90; 0.35]
MI	30	3.53 (0.00)	3.18 (0.00)	0.46*	[0.02; 0.90]
MA	83	3.14 (0.21)	3.53 (0.17)	-0.53	[-1.27; 0.21]
Anforderungsniveau					
EN	185	3.34 (0.16)	3.49 (0.13)	-0.35*	[-0.61; -0.09]
GN	127	3.29 (0.15)	3.35 (0.15)	-0.08	[-0.30; 0.13]
insgesamt	**312**	**3.32 (0.16)**	**3.45 (0.14)**	**-0.23***	**[-0.41; -0.06]**

sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Deutsch dargestellt.

Diese Metaanalysen wurden für die folgenden Fragebogenitems durchgeführt:

- Der Schwierigkeitsgrad der Aufgabe ist insgesamt… (5-stufig, deutlich zu niedrig … deutlich zu hoch)

- Die Bewertungshinweise ermöglichen eine angemessene Einschätzung der Prüfungsleistung. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Die Aufgabenstellung ist klar und verständlich formuliert. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Für die Bearbeitung der Aufgabe ist domänenspezifisches Wissen erforderlich. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Der Text bzw. die Texte der Aufgabe sind angemessen im Hinblick auf die sprachliche Komplexität. (4-stufig, trifft gar nicht zu … trifft voll zu)

Im Rahmen der Metaanalysen wurden die Unterschiede zwischen den Lehrkräfteeinschätzungen zu den Poolaufgaben und den landeseigenen Aufgaben getrennt nach Anforderungsniveau und insgesamt betrachtet. Zusätzlich wurde nach Aufgabenarten unterschieden.

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 (zum Beispiel) der Schwierigkeitsgrad der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als niedriger beurteilt wurde. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte (zum Beispiel) den Schwierigkeitsgrad der Poolaufgaben höher einschätzten als bei den landeseigenen Aufgaben. Für das berechnete Effektmaß ist zusätzlich die untere und obere Grenze des Vertrauensbereichs[2] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 2: Ergebnisse zum Fragebogenitem „Der Schwierigkeitsgrad der Aufgabe ist insgesamt…" (5-stufig, deutlich zu niedrig … deutlich zu hoch)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges' g*)	Vertrauensbereich (95%)
Aufgabenart[3]					
IL	285	3.25 (0.11)	3.28 (0.13)	0.04	[-0.35; 0.44]
EP	179	3.30 (0.22)	3.33 (0.14)	0.30	[-0.47; 1.08]
AP	116	3.34 (0.26)	3.19 (0.10)	0.17	[-0.84; 1.17]
EL	145	3.21 (0.17)	3.09 (0.10)	0.41	[-0.16; 0.97]
MI	32	2.88 (0.00)	3.99 (0.00	-1.52*	[-1.99; -1.05]
MA	84	3.60 (0.23)	3.18 (0.20)	0.61	[-0.34; 1.56]

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

[3] IL = Interpretation literarischer Texte, EP = Erörterung pragmatischer Texte, AP = Analyse pragmatischer Texte, EL = Erörterung literarischer Texte, MI = Materialgestütztes Verfassen informierender Texte, MA = Materialgestütztes Verfassen argumentierender Texte.

Anforderungsniveau[4]					
EN	216	3.31 (0.16)	3.28 (0.16)	0.36	[-0.03; 0.76]
GN	161	3.28 (0.25)	3.26 (0.12)	-0.01	[-0.28; 0.26]
insgesamt	**377**	**3.30 (0.20)**	**3.27 (0.14)**	**0.20**	**[-0.05; 0.46]**

[4] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

Tabelle 3: Ergebnisse zum Fragebogenitem „Die Bewertungshinweise ermöglichen eine angemessene Einschätzung der Prüfungsleistung.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Aufgabenart[5]					
IL	280	3.11 (0.17)	3.07 (0.10)	0.07	[-0.14; 0.27]
EP	120	2.81 (0.14)	2.94 (0.11)	-0.15	[-0.51; 0.21]
AP	118	3.13 (0.19)	3.10 (0.08)	0.09	[-0.19; 0.37]
EL	85	3.09 (0.18)	3.15 (0.10)	-0.10	[-0.71; 0.50]
MI	30	3.13 (0.00)	2.66 (0.00)	0.63*	[0.19; 1.08]
MA	85	2.91 (0.08)	3.16 (0.08)	-0.38*	[-0.60; -0.15]
Anforderungsniveau[6]					
EN	188	3.08 (0.16)	3.08 (0.09)	-0.06	[-0.24; 0.12]
GN	131	2.96 (0.15)	3.00 (0.11)	0.01	[-0.15; 0.18]
insgesamt	**319**	**3.04 (0.17)**	**3.05 (0.10)**	**-0.02**	**[-0.15; 0.10]**

Tabelle 4: Ergebnisse zum Fragebogenitem „Die Aufgabenstellung ist klar und verständlich formuliert.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	282	3.67 (0.15)	3.48 (0.14)	0.27	[-0.00; 0.55]
EP	120	3.47 (0.15)	3.47 (0.14)	-0,01	[-0.30; 0.27]
AP	116	3.65 (0.12)	3.38 (0.17)	0,44	[-0.22; 0.11]
EL	86	3.33 (0.13)	3.36 (0.17)	-0,11	[-0.71; 0.48]
MI	30	3.57 (0.00)	3.30 (0.00)	0,39	[-0.04; 0.83]
MA	84	3.48 (0.11)	3.60 (0.17)	-0,23	[-0.67; 0.21]
Anforderungsniveau					
EN	188	3.57 (0.16)	3.54 (0.11)	0.03	[-0.15; 0.21]
GN	130	3.55 (0.15)	3.39 (0.21)	0.32*	[0.02; 0.61]
insgesamt	**318**	**3.56 (0.16)**	**3.46 (0.15)**	**0.13**	**[-0.03; 0.29]**

[5] IL = Interpretation literarischer Texte, EP = Erörterung pragmatischer Texte, AP = Analyse pragmatischer Texte, EL = Erörterung literarischer Texte, MI = Materialgestütztes Verfassen informierender Texte, MA = Materialgestütztes Verfassen argumentierender Texte.

[6] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

Tabelle 5: Ergebnisse zum Fragebogenitem „Für die Bearbeitung der Aufgabe ist domänenspezifisches Wissen erforderlich.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Aufgabenart[7]					
IL	277	3.45 (0.11)	3.49 (0.10)	-0.05	[-0.22; 0.13]
EP	121	3.12 (0.13)	3.48 (0.15)	-0.53*	[-0.84; -0.22]
AP	111	3.41 (0.09)	3.50 (0.05)	-0.15	[-0.62; 0.33]
EL	83	3.59 (0.11)	3.38 (0.17)	0.30	[-0.06; 0.66]
MI	30	3.10 (0.00)	3.54 (0.00)	-0.63*	[-1.07; -0.18]
MA	84	3.46 (0.16)	3.29 (0.22)	-0.02	[-1.13; 1.10]
Anforderungsniveau[8]					
EN	184	3.45 (0.14)	3.42 (0.18)	-0.07	[-0.30; 0.16]
GN	128	3.28 (0.14)	3.38 (0.11)	-0.18	[-0.40; 0.03]
insgesamt	**312**	**3.39 (0.14)**	**3.45 (0.14)**	**-0.11**	**[-0.27; 0.05]**

Tabelle 6: Ergebnisse zum Fragebogenitem „Der Text bzw. die Texte der Aufgabe sind angemessen im Hinblick auf die sprachliche Komplexität.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Aufgabenart					
IL	278	3.46 (0.12)	3.46 (0.13)	0.00	[-0.24; 0.23]
EP	120	3.11 (0.16)	3.41 (0.19)	-0.56	[-1.21; 0.08]
AP	112	3.21 (0.13)	3.46 (0.08)	-0.30	[-0.66; 0.05]
EL	83	3.39 (0.17)	3.45 (0.09)	-0.28	[-0.90; 0.35]
MI	30	3.53 (0.00)	3.18 (0.00)	0.46*	[0.02; 0.90]
MA	83	3.14 (0.21)	3.53 (0.17)	-0.53	[-1.27; 0.21]
Anforderungsniveau					
EN	185	3.34 (0.16)	3.49 (0.13)	-0.35*	[-0.61; -0.09]
GN	127	3.29 (0.15)	3.35 (0.15)	-0.08	[-0.30; 0.13]
insgesamt	**312**	**3.32 (0.16)**	**3.45 (0.14)**	**-0.23***	**[-0.41; -0.06]**

[7] IL = Interpretation literarischer Texte, EP = Erörterung pragmatischer Texte, AP = Analyse pragmatischer Texte, EL = Erörterung literarischer Texte, MI = Materialgestütztes Verfassen informierender Texte, MA = Materialgestütztes Verfassen argumentierender Texte.

[8] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

Der Schwierigkeitsgrad der Poolaufgaben im Fach Deutsch wurde aufgabenübergreifend mehrheitlich als „angemessen“ bewertet. Außerdem wurde in einer Metaanalyse der Schwierigkeitseinschätzungen der Lehrkräfte aufgabenübergreifend keine statistisch signifikante Differenz zwischen Poolaufgaben und landeseigenen Aufgaben ermittelt (g = 0.20)[9]. Auch bei einer nach Aufgabenarten differenzierten Analyse lassen sich in den meisten Fällen keine signifikanten Unterschiede zwischen den Schwierigkeitseinschätzungen zu den aus dem Pool entnommenen Aufgaben und den landeseigenen Aufgaben feststellen. Die einzige Ausnahme hiervon bilden die Poolaufgaben zum Materialgestützten Verfassen informierender Texte (MI, signifikant leichter eingeschätzt, g = -1.52, Differenz als „hoch“ einzustufen).

Der Aussage, dass die Bewertungshinweise eine angemessene Einschätzung der Prüfungsleistung ermöglichen, stimmten die befragten Lehrkräfte im Mittel „eher zu“, wobei die Poolaufgaben und die landeseigenen Aufgaben im Hinblick auf diesen Aspekt aufgabenübergreifend betrachtet sehr ähnlich beurteilt wurden (g = -0.02). Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben fanden sich bei einer nach Aufgabenarten differenzierten Betrachtung nur für die Aufgabenarten MI (signifikant höhere Zustimmung bei den Poolaufgaben, g = 0.63, als „mittel“ einzustufen) und Materialgestütztes Verfassen argumentierender Texte (MA, signifikant geringere Zustimmung bei den Poolaufgaben, g = -0.38, als „gering“ einzustufen).

Sehr positiv wurde die Formulierung der Aufgabenstellungen beurteilt, die aus Sicht der allermeisten Lehrkräfte klar und verständlich war. Im Hinblick auf diesen Aspekt wurden die Poolaufgaben insgesamt etwas positiver als die landeseigenen Aufgaben beurteilt (g = 0.13, nicht statistisch signifikant), wobei dieser Unterschied nur für die Aufgaben auf grundlegendem Niveau statistisch signifikant ausfällt (g = 0.32, als „gering“ einzustufen).

Der Aussage, dass für die Bearbeitung der Aufgabe domänenspezifisches Wissen erforderlich sei, stimmten die befragten Lehrkräfte im Mittel „eher [bis] voll zu“, wobei die Poolaufgaben und die landeseigenen Aufgaben in Bezug auf diesen Aspekt aufgabenübergreifend betrachtet ähnlich eingeschätzt wurden (g = -0.11, nicht statistisch signifikant). Bei einem nach Aufgabenarten differenzierten Vergleich zwischen Poolaufgaben und landeseigenen Aufgaben finden sich signifikant geringere Zustimmungswerte für die Poolaufgaben der Art Erörterung pragmatischer Texte (EP, g = -0.53, als „mittel“ einzustufen) und MI (g = -0.63, als „mittel“ einzustufen).

Die sprachliche Komplexität der den Poolaufgaben zugrundeliegenden Texte wurde von den befragten Lehrkräften zumeist als „angemessen“ oder „eher angemessen“ eingestuft. Dabei wurde dieser Aspekt aufgabenübergreifend betrachtet bei den Poolaufgaben etwas weniger positiv eingeschätzt als bei den landeseigenen Aufgaben (g = -0.23, als „gering“ einzustufen). Eine nach Anforderungsniveaus differenzierte Betrachtung zeigt, dass die Unterschiede vor allem die Aufgaben zum erhöhten Niveau betreffen (g = -0.35, statistisch signifikant und als „gering“ einzustufen). Für die Aufgaben zum grundlegenden Niveau fällt die Differenz deutlich geringer und statistisch nicht signifikant aus (g = -0.08.)

[9] Positive/Negative Differenz = Poolaufgaben werden als schwieriger/als weniger schwierig eingeschätzt als die landeseigenen Aufgaben.

2.2 Englisch

In den folgenden Tabellen sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Englisch dargestellt, wobei nur die Einschätzungen zu Aufgaben der beiden Domänen Schreiben und Sprachmittlung berücksichtigt wurden.

Diese Metaanalysen wurden für die folgenden Fragebogenitems durchgeführt:

- Der Schwierigkeitsgrad der Aufgabe ist insgesamt… (5-stufig, deutlich zu niedrig … deutlich zu hoch)
- Die Bewertungshinweise ermöglichen eine angemessene Einschätzung der Prüfungsleistung. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Die Aufgabenstellung ist klar und verständlich formuliert. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Der zugrundeliegende Text ist angemessen im Hinblick auf die sprachliche Komplexität … (4-stufig, trifft gar nicht zu … trifft voll zu)

Im Rahmen der Metaanalysen wurden die Unterschiede zwischen den Lehrkräfteeinschätzungen zu den Poolaufgaben und den landeseigenen Aufgaben getrennt nach Anforderungsniveau und insgesamt betrachtet. Zusätzlich wurde nach den beiden Domänen Schreiben und Sprachmittlung unterschieden.

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 (zum Beispiel) der Schwierigkeitsgrad der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als niedriger beurteilt wurde. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte (zum Beispiel) den Schwierigkeitsgrad der Poolaufgaben höher einschätzten als bei den landeseigenen Aufgaben. Für das berechnete Effektmaß ist zusätzlich die untere und obere Grenze des Vertrauensbereichs[10] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

[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.

Tabelle 7: Ergebnisse zum Fragebogenitem „Der Schwierigkeitsgrad der Aufgabe ist insgesamt…“ (5-stufig, deutlich zu niedrig … deutlich zu hoch)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Kompetenzbereich[11]					
S	304	3.10 (0.15)	3.25 (0.13)	-0.08	[-0.30; 0.15]
SM	355	3.10 (0.11)	3.20 (0.14)	-0.20	[-0.43; 0.04]
Anforderungsniveau[12]					
EN	256	3.09 (0.11)	3.19 (0.14)	-0.12	[-0.34; 0.10]
GN	132	3.13 (0.17)	3.25 (0.13)	-0.17	[-0.41; 0.07]
insgesamt	**388**	**3.10 (0.13)**	**3.22 (0.13)**	**-0.15**	**[-0.30; 0.01]**

Tabelle 8: Ergebnisse zum Fragebogenitem „Die Bewertungshinweise ermöglichen eine angemessene Einschätzung der Prüfungsleistung.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Kompetenzbereich					
S	243	3.12 (0.07)	3.17 (0.07)	-0.24	[-0.49; 0.00]
SM	270	3.33 (0.10)	3.22 (0.07)	0.14	[-0.03; 0.30]
Anforderungsniveau					
EN	223	3.18 (0.08)	3.20 (0.08)	-0.10	[-0.31; 0.11]
GN	94	3.33 (0.10)	3.17 (0.08)	0.14	[-0.15; 0.43]
insgesamt	**317**	**3.22 (0.09)**	**3.20 (0.07)**	**-0.02**	**[-0.18; 0.14]**

[11] S = Schreiben, SM = Sprachmittlung.
[12] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

Tabelle 9: Ergebnisse zum Fragebogenitem „Die Aufgabenstellungen sind klar und verständlich formuliert.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Kompetenzbereich[13]					
S	243	3.48 (0.15)	3.56 (0.10)	-0.26	[-0.65; 0.14]
SM	295	3.49 (0.20)	3.55 (0.07)	-0.09	[-0.35; 0.17]
Anforderungsniveau[14]					
EN	230	3.48 (0.17)	3.55 (0.10)	-0.18	[-0.46; 0.10]
GN	98	3.49 (0.18)	3.46 (0.04)	-0.08	[-0.44; 0.29]
insgesamt	**328**	**3.49 (0.18)**	**3.56 (0.08)**	**-0.15**	**[-0.35; 0.05]**

Tabelle 10: Ergebnisse zum Fragebogenitem „Der zugrundeliegende Text ist angemessen im Hinblick auf die sprachliche Komplexität.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Kompetenzbereich					
S	239	3.58 (0.10)	3.43 (0.14)	0.11	[-0.14; 0.36]
SM	292	3.51 (0.10)	3.54 (0.11)	-0.08	[-0.28; 0.11]
Anforderungsniveau					
EN	225	3.56 (0.10)	3.52 (0.14)	0.05	[-0.11; 0.22]
GN	97	3.52 (0.09)	3.47 (0.10)	-0.14	[-0.46; 0.18]
insgesamt	**322**	**3.55 (0.10)**	**3.49 (0.13)**	**-0.01**	**[-0.15; 0.13]**

Im Fach Englisch sind die Einschätzungen der Lehrkräfte zu den Poolaufgaben überwiegend positiv. Der Schwierigkeitsgrad der Poolaufgaben wurde überwiegend als angemessen beurteilt, auch bestätigten die meisten Lehrkräfte, dass die Bewertungshinweise eine angemessene Einschätzung der Prüfungsleistungen ermöglichen, die Aufgabenstellungen der Poolaufgaben klar und verständlich formuliert sind und die den Poolaufgaben zugrundeliegenden Texte im Hinblick auf die sprachliche Komplexität angemessen sind. Um aufgabenübergreifende Aussagen zu ermöglichen, wurden auch im Fach Englisch Metanalysen zu den genannten Fragebogeninhalten durchgeführt. Hierbei wurden weder in der Gesamtbetrachtung, noch bei einer Differenzierung nach Anforderungsniveaus und Domänen, signifikante Unterschiede zwischen den Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben festgestellt.

[13] S = Schreiben, SM = Sprachmittlung.
[14] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

2.3 Französisch

Im Fach Französisch wurde die Lehrkräftebefragung (auf der Grundlage von Vorabinformationen zur Anzahl der Prüfungen in diesem Fach und nach Abstimmung mit den Ländern) nur in sechs Ländern durchgeführt. Ein aussagekräftiger Stichprobenumfang wurde dabei nur in zwei Ländern erreicht, in den vier übrigen Ländern beteiligten sich nur vereinzelt Lehrkräfte an der Befragung. Aufgrund der geringen Stichprobenumfänge musste auf eine aufgabenübergreifende metanalytische Auswertung verzichtet werden. Die wenigen vorliegenden Lehrkräfteeinschätzungen zu den Poolaufgaben weisen aber auf ein ähnlich positives Bild wie im Fach Englisch hin.

2.4 Mathematik

In den folgenden Tabelle 11Tabellen sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Mathematik dargestellt, getrennt nach den Prüfungsteilen A (hilfsmittelfrei) und B (mit Hilfsmittel WTR bzw. MMS). Diese Metaanalysen wurden für die folgenden Fragebogenitems durchgeführt:

- Der Schwierigkeitsgrad der Aufgabe ist insgesamt… (5-stufig, deutlich zu niedrig … deutlich zu hoch)
- Die Dichte der Aufgaben ist angemessen. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Die Aufgabenstellungen sind hinsichtlich ihrer sprachlichen Komplexität der Prüfung angemessen. (4-stufig, trifft gar nicht zu … trifft voll zu)
- Die Aufgabenstellungen sind hinsichtlich ihrer sprachlichen Komplexität schülerge-recht formuliert. (4-stufig, trifft gar nicht zu … trifft voll zu)

Für den Prüfungsteil A wurden die Unterschiede zwischen den Lehrkräfteeinschätzungen zu den Poolaufgaben und den landeseigenen Aufgaben getrennt nach Anforderungsniveau und insgesamt betrachtet, für den Prüfungsteil B wurde zusätzlich nach Sachgebieten unterschieden.

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 (zum Beispiel) der Schwierigkeitsgrad der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als niedriger beurteilt wurde. Demgegenüber gibt ein positives Vorzeichen an, dass die Lehrkräfte (zum Beispiel) den Schwierigkeitsgrad der Poolaufgaben höher einschätzten als bei den landeseigenen Aufgaben. Für das berechnete Effektmaß ist zusätzlich die untere und obere Grenze des Vertrauensbereichs[15] angegeben. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

[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.

Tabelle 11: Ergebnisse zum Prüfungsteil A (hilfsmittelfrei), Fragebogenitem „Der Schwierigkeitsgrad der Aufgabe ist insgesamt…“ (5-stufig, deutlich zu niedrig … deutlich zu hoch)

Anforderungsniveau[16]	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
EN	248	3.36 (0.15)	3.47 (0.05)	-0.17	[-0.39; 0.04]
GN	94	3.33 (0.14)	3.10 (0.09)	0.31	[-3.51; 4.13]
insgesamt	**342**	**3.36 (0.15)**	**3.43 (0.06)**	**-0.13**	**[-0.34; 0.08]**

Tabelle 12: Ergebnisse zum Prüfungsteil B, Fragebogenitem „Der Schwierigkeitsgrad der Aufgabe ist insgesamt…“ (5-stufig, deutlich zu niedrig … deutlich zu hoch)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Sachgebiet[17]					
A	381	3.63 (0.13)	3.36 (0.08)	0.49*	[0.30; 0.68]
AG	284	3.98 (0.10)	3.56 (0.06)	0.57*	[0.24; 0.89]
S	266	3.43 (0.17)	3.38 (0.09)	0.24	[-0.11; 0.68]
Anforderungsniveau					
EN	290	3.66 (0.11)	3.48 (0.07)	0.36*	[0.14; 0.58]
GN	147	3.63 (0.16)	3.33 (0.09)	0.50*	[0.30; 0.71]
insgesamt	**437**	**3.64 (0.14)**	**3.40 (0.08)**	**0.43***	**[0.29; 0.58]**

Tabelle 13: Ergebnisse zum Prüfungsteil A (hilfsmittelfrei), Fragebogenitem „Die Dichte der Aufgaben ist angemessen.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

Anforderungsniveau	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
EN	241	3.16 (0.12)	2.80 (0.08)	0.46*	[0.25; 0.66]
GN	93	3.10 (0.10)	3.13 (0.03)	0.04	[-4.19; 4.28]
insgesamt	**334**	**3.15 (0.12)**	**2.84 (0.09)**	**0.42***	**[0.22; 0.62]**

[16] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.
[17] A = Analysis, AG = Analytische Geometrie, S = Stochastik.

Tabelle 14: Ergebnisse zum Prüfungsteil B, Fragebogenitem „Die Dichte der Aufgaben ist angemessen.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Sachgebiet[18]					
A	331	2.67 (0.13)	2.88 (0.08)	-0.28*	[-0.45; -0.12]
AG	221	2.54 (0.13)	2.62 (0.06)	-0.11	[-0.31; 0.10]
S	220	3.14 (0.15)	2.86 (0.12)	0.27*	[0.02; 0.53]
Anforderungsniveau[19]					
EN	242	2.80 (0.12)	2.73 (0.10)	0.04	[-0.18; 0.25]
GN	143	2.72 (0.16)	2.91 (0.06)	-0.24*	[-0.38; -0.09]
insgesamt	**385**	**2.76 (0.14)**	**2.83 (0.09)**	**-0.11**	**[-0.24; 0.02]**

Tabelle 15: Ergebnisse zum Prüfungsteil A (hilfsmittelfrei), Fragebogenitem „Die Aufgabenstellungen sind hinsichtlich ihrer sprachlichen Komplexität der Prüfung angemessen.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

Anforderungsniveau	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
EN	240	3.57 (0.12)	3.23 (0.10)	0.44*	[0.24; 0.65]
GN	92	3.33 (0.12)	3.50 (0.09)	-0.09	[-4.70; 4.52]
insgesamt	**332**	**3.55 (0.15)**	**3.26 (0.10)**	**0.39***	**[0.19; 0.59]**

[18] A = Analysis, AG = Analytische Geometrie, S = Stochastik.
[19] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

Tabelle 16: Ergebnisse zum Prüfungsteil B, Fragebogenitem „Die Aufgabenstellungen sind hinsichtlich ihrer sprachlichen Komplexität der Prüfung angemessen.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Sachgebiet[20]					
A	321	2.95 (0.11)	3.24 (0.09)	-0.45*	[-0.60; -0.30]
AG	272	2.87 (0.12)	3.25 (0.08)	-0.50*	[-0.72; -0.28]
S	213	3.16 (0.16)	3.30 (0.09)	-0.29*	[-0.58; -0.00]
Anforderungsniveau[21]					
EN	231	2.98 (0.13)	3.25 (0.09)	-0.41*	[-0.59; -0.23]
GN	141	2.98 (0.12)	3.26 (0.09)	-0.41*	[-0.56; -0.27]
insgesamt	**372**	**2.98 (0.13)**	**3.26 (0.09)**	**-0.41***	**[-0.52; -0.30]**

Tabelle 17: Ergebnisse zum Prüfungsteil A (hilfsmittelfrei), Fragebogenitem „Die Aufgabenstellungen sind hinsichtlich ihrer sprachlichen Komplexität schülergerecht formuliert.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

Anforderungsniveau	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
EN	240	3.42 (0.18)	3.00 (0.15)	0.47*	[0.29; 0.64]
GN	93	3.17 (0.17)	3.39 (0.13)	-0.11	[-5.54; 5.32]
insgesamt	**333**	**3.39 (0.18)**	**3.05 (0.16)**	**0.40***	**[0.22; 0.59]**

[20] A = Analysis, AG = Analytische Geometrie, S = Stochastik.
[21] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.

Tabelle 18: Ergebnisse zum Prüfungsteil B, Fragebogenitem „Die Aufgabenstellungen sind hinsichtlich ihrer sprachlichen Komplexität schülergerecht formuliert.“ (4-stufig, trifft gar nicht zu … trifft voll zu)

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Sachgebiet[22]					
A	321	2.72 (0.14)	3.05 (0.12)	-0.47*	[-0.65; -0.28]
AG	274	2.76 (0.15)	3.03 (0.15)	-0.28*	[-0.48; -0.08]
S	216	2.98 (0.17)	3.06 (0.15)	-0.21	[-0.49; 0.07]
Anforderungsniveau[23]					
EN	233	2.84 (0.17)	2.99 (0.13)	-0.24*	[-0.42; -0.05]
GN	141	2.75 (0.13)	3.09 (0.13)	-0.45*	[-0.61; -0.30]
insgesamt	**374**	**2.79 (0.15)**	**3.05 (0.13)**	**-0.36***	**[-0.48; -0.24]**

Im Fach Mathematik ergeben die Ergebnisse der Lehrkräftebefragung zum Schwierigkeitsgrad der Poolaufgaben ein heterogenes Bild. Im hilfsmittelfreien Prüfungsteil A wurde der Schwierigkeitsgrad der Poolaufgaben im Mittel als „angemessen“ bis „etwas zu hoch“ eingestuft, zwischen Poolaufgaben und landeseigenen Aufgaben wurden in Bezug auf die Einschätzungen dieses Aspekts keine statistisch signifikanten Unterschiede ermittelt (g = -0.13)[24]. Im Prüfungsteil B, in dem bei der Bearbeitung von Aufgaben Hilfsmittel zugelassen sind, wurde der Schwierigkeitsgrad der Poolaufgaben überwiegend als „eher zu hoch“ beurteilt. Aufgabenübergreifend betrachtet wurde der Schwierigkeitsgrad der Poolaufgaben signifikant höher eingeschätzt als der Schwierigkeitsgrad der landeseigenen Aufgaben (g = 0.43, als „gering“ bis „mittel“ einzustufen). Diese Differenz fällt für die Aufgaben des grundlegenden Niveaus im Prüfungsteil B etwas höher aus (g = 0.50, als „mittel“ einzustufen) als für die Aufgaben des erhöhten Niveaus (g = 0.36, als „gering“ bis „mittel“ einzustufen).

Die Einschätzungen zur Dichte der Aufgaben fallen ebenfalls je nach Prüfungsteil unterschiedlich aus. Im Prüfungsteil A stimmten die Lehrkräfte der Aussage, dass die Dichte der Aufgaben angemessen sei, im Mittel „eher zu“, wobei die Dichte der Poolaufgaben positiver eingeschätzt wurde als die Dichte der landeseigenen Aufgaben (g = 0.42, statistisch signifikant, als „gering“ bis „mittel“ einzustufen). Im Prüfungsteil B sind die Zustimmungswerte insbesondere bei den Poolaufgaben geringer als im Prüfungsteil A, wobei die Dichte der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben nur geringfügig als weniger angemessen eingestuft wurde (g = -0.11, nicht statistisch signifikant). Im Prüfungsteil B findet sich nur für die Aufgaben des grundlegenden Niveaus ein signifikanter Unterschied zwischen Poolaufgaben und landeseigene Aufgaben (g = -0.24, als „gering“ einzustufen).

Im Rahmen der Befragung wurden die teilnehmenden Lehrkräfte außerdem um Einschätzungen zur Formulierung der Aufgabenstellungen gebeten. Unter anderem sollte beurteilt werden, ob die Formulierungen schülergerecht gewählt und hinsichtlich der sprachlichen Komplexität

[22] A = Analysis, AG = Analytische Geometrie, S = Stochastik.
[23] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.
[24] Positive/Negative Differenz = Poolaufgaben werden als schwieriger/als weniger schwierig eingeschätzt als die landeseigenen Aufgaben.

der Prüfung angemessen sind. Auch im Hinblick auf diese beiden Aspekte fallen die Lehrkräfteeinschätzungen je nach Prüfungsteil unterschiedlich aus. Im Prüfungsteil A stimmten die Lehrkräfte den beiden Aussagen zumeist „eher zu“ oder „voll zu“; im Prüfungsteil B fallen die Zustimmungsraten hingegen geringer aus. Im Prüfungsteil A wird die Formulierung der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben im Hinblick auf die abgefragten Aspekte als angemessener eingestuft (sprachliche Komplexität: $g = 0.39$[25], statistisch signifikant, als „gering“ bis „mittel“ einzustufen; schülergerechte Formulierung: $g = 0.40$, statistisch signifikant, als „gering“ bis „mittel“ einzustufen). Im Prüfungsteil B findet sich hingegen ein gegenteiliges Befundmuster, hier wird die Formulierung der Poolaufgaben im Vergleich zu den landeseigenen Aufgaben als weniger angemessen beurteilt (sprachliche Komplexität: $g = -0.41$, statistisch signifikant, als „gering“ bis „mittel“ einzustufen; schülergerechte Formulierung: $g = -0.36$, statistisch signifikant, als „gering“ bis „mittel“ einzustufen).

3 Erhebung der Vor- und Prüfungsleistungen von Schülerinnen und Schülern im Fach Mathematik

Im Rahmen der Evaluation des Einsatzes von Aufgaben der Pools für das Prüfungsjahr 2023 wurden Daten zu den Vor- und Prüfungsleistungen von Schülerinnen und Schülern im Fach Mathematik erhoben, d. h. die Halbjahresnoten und die Klausurnoten der Qualifikationsphase. Zudem wurde erfasst, welche Ergebnisse die Schülerinnen und Schüler der für die Evaluation ausgewählten Kurse bei den einzelnen Teilaufgaben der Prüfung sowie bei der Prüfung insgesamt erzielt haben.

3.1 Empirische Schwierigkeit der Poolaufgaben

Als Indikator für die empirische Schwierigkeit der aus dem Pool für das Fach Mathematik eingesetzten Aufgaben wurde jeweils der arithmetische Mittelwert der von den Prüflingen erzielten Notenpunkte berechnet. Zudem wurde untersucht, ob sich die eingesetzten Aufgaben aus dem Pool hinsichtlich der empirischen Schwierigkeit von den landeseigenen Aufgaben unterscheiden.

In den nachfolgenden Tabellen sind die Ergebnisse der Metaanalysen zur empirischen Schwierigkeit der Teilaufgaben für das Fach Mathematik dargestellt. Hierfür wurden für jedes Land die Differenzen zwischen der mittleren empirischen Schwierigkeit der Teilaufgaben aus dem Pool und der mittleren empirischen Schwierigkeit der landeseigenen Teilaufgaben ermittelt, getrennt nach den Prüfungsteilen A (hilfsmittelfrei) und B (mit Hilfsmittel WTR bzw. MMS). Diese Differenzen wurden standardisiert und über alle Länder aggregiert, die an der Evaluation teilgenommen haben. Für den Prüfungsteil A wurden die Unterschiede in den Lösungsquoten bei Teilaufgaben aus dem Pool und landeseigenen Teilaufgaben getrennt nach Anforderungsniveau und insgesamt betrachtet, für den Prüfungsteil B wurde zusätzlich nach Sachgebieten unterschieden. 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

[25] Positive/Negative Differenz = Poolaufgaben werden als schwieriger/als weniger schwierig eingeschätzt als die landeseigenen Aufgaben.

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[26] sind in den folgenden Tabellen dargestellt. Eine negative Differenz bedeutet, dass Prüflinge bei Aufgaben aus dem Pool im Durchschnitt ein weniger gutes Ergebnis erzielten als Prüflinge, die landeseigene Aufgaben wählten. Eine positive Differenz weist darauf hin, dass Prüflinge bei der Bearbeitung von Aufgaben aus dem Pool bessere Ergebnisse erzielten. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 19: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Prüfungsteil A (hilfsmittelfrei) im Fach Mathematik

Anfordrungs-niveau[27]	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
EN	4115	0,58 (0,04)	0,67 (0,05)	-0,37*	[-0,61; -0,14]
GN	2067	0,57 (0,06)	0,60 (0,02)	-0,08	[-0,28; 0,11]
insgesamt	**6182**	**0,58 (0,05)**	**0,65 (0,04)**	**-0,25***	**[-0,41; -0,09]**

Tabelle 20: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Prüfungsteil B im Fach Mathematik

	N	MW (SD) (Pool)	MW (SD) (Land)	Stand. Differenz (*Hedges‘ g*)	Vertrauensbereich (95%)
Sachgebiet[28]					
A	5985	0,53 (0,03)	0,58 (0,03)	-0,28*	[-0,52; -0,05]
AG	5103	0,57 (0,20)	0,53 (0,05)	0,10	[-0,42; 0,61]
S	3631	0,58 (0,03)	0,61 (0,05)	-0,04	[-0,57; 0,49]
Anforderungsniveau					
EN	3919	0,57 (0,12)	0,59 (0,04)	-0,10	[-0,33; 0,12]
GN	2066	0,49 (0,02)	0,55 (0,03)	-0,23	[-0,62; 0,16]
insgesamt	**5985**	**0,54 (0,10)**	**0,57 (0,04)**	**-0,15**	**[-0,34; 0,04]**

Die Ergebnisse der Metanalyse lassen sich wie folgt zusammenfassen:

- Für den Prüfungsteil A lassen die Ergebnisse der Metaanalyse darauf schließen, dass die Prüflinge bei den Poolaufgaben etwas weniger gute Leistungen erzielten als bei den

[26] Als Vertrauensbereich wird jeweils das auf der Grundlage der Stichproben ermittelte Intervall bezeichnet, das den tatsächlichen Wert mit einer Wahrscheinlichkeit von 95 Prozent enthält.
[27] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau.
[28] A = Analysis, AG = Analytische Geometrie, S = Stochastik.

landeseigenen Aufgaben (g = -0.25, statistisch signifikant), wobei die Größe des ermittelten Unterschieds als „gering“ einzustufen ist. Eine nach Anforderungsniveaus differenzierte Analyse zeigt, dass dieser Unterschied nur für die Aufgaben des erhöhten Niveaus (g = -0.37, als „gering“ bis „mittel“ einzustufen), nicht jedoch für die Aufgaben des grundlegenden Niveaus (g = -0.08) statistisch signifikant ausfällt.

- Für den Prüfungsteil B weisen die Ergebnisse der Metaanalysen zwar deskriptiv darauf hin, dass die Prüflinge bei den Poolaufgaben insgesamt etwas weniger gut abgeschnitten haben als bei den landeseigenen Aufgaben, die ermittelte Differenz ist jedoch gering und statistisch nicht signifikant (g = -0.15). Auch bei einer separaten Betrachtung des erhöhten Niveaus (g = -0.10) und des grundlegenden Niveaus (g = -0.23) wurde kein statistisch signifikanter Unterschied zwischen den bei den Poolaufgaben und bei den landeseigenen Aufgaben erreichten Lösungsquoten ermittelt. Bei einer nach Sachgebieten differenzierten Analyse zeigt sich, dass die Prüflinge bei den Poolaufgaben zur Analysis signifikant geringere Leistungen erzielten als bei den landeseigenen Aufgaben zu diesem Sachgebiet (g = -0.28, als „gering“ einzustufen). Für die übrigen Sachgebiete fanden sich keine signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben.

Insgesamt ist festzuhalten, dass sich im Rahmen der Evaluation des Einsatzes von Aufgaben des Pools für das Prüfungsjahr 2023 – anders noch als bei der Evaluation zum Prüfungsjahr 2021 – zwar zeigte, dass die Prüflinge bei den Poolaufgaben tendenziell etwas weniger gut abgeschnitten haben als bei den landeseigenen Aufgaben. Die festgestellten Unterschiede sind jedoch nicht in jedem Fall statistisch signifikant und fallen vor allem beim Prüfungsteil B deutlich geringer aus als es die Ergebnisse der Lehrkräftebefragung vermuten lassen.

3.2 Kriteriale Validität der Poolaufgaben

Als Indikatoren für die kriteriale Validität 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[29] sind in den beiden folgenden Tabellen getrennt 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.

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

Tabelle 21: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Mathematik

Sachgebiete[30]	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben		
HF	0.74	[0.70; 0.77]
A	0.73	[0.69; 0.76]
AG	0.69	[0.61; 0.75]
S	0.72	[0.64; 0.78]
insgesamt	**0.78**	**[0.74; 0.81]**
landeseigene Aufgaben		
HF	0.65	[0.59; 0.70]
A	0.74	[0.70; 0.77]
AG	0.71	[0.64; 0.76]
S	0.70	[0.64; 0.74]
insgesamt	**0.77**	**[0.75; 0.80]**

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

Sachgebiete	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben		
HF	0.73	[0.69; 0.76]
A	0.72	[0.68; 0.76]
AG	0.67	[0.60; 0.73]
S	0.69	[0.61; 0.76]
insgesamt	**0.77**	**[0.73; 0.80]**
landeseigene Aufgaben		
HF	0.65	[0.59; 0.70]
A	0.72	[0.67; 0.77]
AG	0.70	[0.64; 0.76]
S	0.70	[0.64; 0.74]
insgesamt	**0.78**	**[0.74; 0.81]**

[30] HF = Hilfsmittelfreier Teil, A = Analysis, AG = Analytische Geometrie, S = Stochastik.

Die als Indikatoren zur kriterialen Validität der Aufgaben im Fach Mathematik ermittelten Ergebnisse lassen sich wie folgt zusammenfassen:

- Der in Bezug auf das Validitätskriterium „Halbjahresnoten“ für die Teilaufgaben aus dem Pool berechnete Validitätskoeffizient ist als hoch einzustufen (r = .78) und unterscheidet sich nicht statistisch signifikant vom entsprechenden Kennwert für die landeseigenen Teilaufgaben (r = .77). Der kleinste Validitätskoeffizient wurde für die Teilaufgaben aus dem Pool zum Sachgebiet „Analytische Geometrie“ ermittelt (r = .69, als „hoch“ einzustufen). Innerhalb der einzelnen Sachgebiete zeigen sich für die Validitätskoeffizienten keine statistisch signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben.
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Teilaufgaben aus dem Pool ein aggregierter Validitätskoeffizient von r = .77 (als „hoch“ einzustufen) ermittelt. Dieser Koeffizient unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde (r = .78, als „hoch“ einzustufen). Der kleinste Validitätskoeffizient wurde wiederum für die Teilaufgaben aus dem Pool zum Sachgebiet „Analytische Geometrie“ ermittelt (r = .67, als „hoch“ einzustufen). Auch für die Klausurnoten ergeben sich innerhalb der einzelnen Sachgebiete für die Validitätskoeffizienten keine statistisch signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben.

Positiv hervorzuheben ist, dass die für die Poolaufgaben zum Sachgebiet „Stochastik“ ermittelten Validitätskoeffizienten deutlich höher ausfallen als in den Prüfungsjahren 2021 und 2017. 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.

4 Literatur

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.