Institut zur Qualitätsentwicklung
im Bildungswesen

Gemeinsame Abituraufgabenpools der Länder

Evaluation von Aufgaben der Pools für das Prüfungsjahr 2022

Ergebnisse zur Bewährung der Aufgaben

Dr. Lars Hoffmann, Anja Riemenschneider, 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 2022 dar. Diese Evaluation umfasst drei Bereiche:

- eine Befragung der Lehrkräfte zu den eingesetzten Poolaufgaben
- eine Befragung von Vertreterinnen und Vertretern der Abiturkommissionen der Länder zu den eingesetzten Poolaufgaben
- eine Erhebung und Auswertung von Ergebnissen zu den Vor- und Prüfungsleistungen der Schülerinnen und Schüler

Befragung der Lehrkräfte

In den Fächern Deutsch, Englisch und Französisch fallen die Rückmeldungen der befragten Lehrkräfte zu den Poolaufgaben in Bezug auf alle abgefragten Aspekte (Schwierigkeitsgrad der Aufgaben, Erwartungshorizonte und Bewertungshinweise zu den Aufgaben, Aufgabenstellungen und Materialien) insgesamt positiv aus. Die wenigen kritischen Rückmeldungen betreffen vor allem die (als zu wenig stark eingeschätzten) Bezüge zu den Lehrplaninhalten und zum Unterrichtswissen. Für das Fach Mathematik ergeben die Ergebnisse der Lehrkräftebefragung ein heterogenes Bild: Alle drei thematisierten Aspekte (Schwierigkeitsgrad der Aufgaben, Aufgabenstellungen, Umfang der Aufgaben) werden in einigen Ländern eher positiv, in anderen Ländern eher negativ eingeschätzt.

Befragung von Vertreterinnen und Vertretern der Abiturkommissionen der Länder

Pro Land und Fach wurde jeweils eine Vertreterin bzw. ein Vertreter um eine Rückmeldung zu den vom jeweiligen Land eingesetzten Poolaufgaben gebeten. In den Fächern Deutsch, Englisch und Französisch fallen diese Rückmeldungen überwiegend positiv aus. In Mathematik zeigen diese Befragungsergebnisse wiederum ein heterogenes Bild, wobei die Rückmeldungen zum Umfang bzw. zur Dichte der Poolaufgaben im Fach Mathematik seitens der Abiturkommissionsmitglieder insgesamt kritischer ausfallen als bei den Lehrkräften.

Ergebnisse zur Erhebung der Vor- und Prüfungsleistungen von Schülerinnen und Schülern

Die in diesem Teil der Evaluation ermittelten Kennwerte zur empirischen Schwierigkeit der Aufgaben und zum statistischen Zusammenhang zwischen den Prüfungsleistungen und den in der Qualifikationsphase erzielten Leistungen (im Folgenden als „kriteriale Validität“ bezeichnet) legen insgesamt den Schluss nahe, dass die Schülerinnen und Schüler bei den Poolaufgaben und den landeseigenen Aufgaben ähnliche Ergebnisse erzielen.

1 Verfahrensbeschreibung

Die Evaluation, deren Ergebnisse im vorliegenden Bericht dokumentiert sind, basiert auf einer mit der AG Abiturkommission abgestimmten Weiterentwicklung des Evaluationskonzepts, dem die Amtschefskommission „Qualitätssicherung in Schulen“ in ihrer 98. Sitzung am 28.01.2021 zugestimmt hat. Der vorliegende Bericht beinhaltet die Ergebnisse zu zwei zentralen Bestandteilen der ersten Säule des Konzepts für das Prüfungsjahr 2022:

(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 2.4 (Deutsch), 3.4 (Englisch), 4.4 (Französisch) und 5 (Mathematik).

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

- Auswahl der Aufgaben – nur für das Fach Deutsch relevant – (Wie häufig werden die Aufgaben aus den Pools im Vergleich zu den landeseigenen Aufgaben gewählt (sofern Wahlmöglichkeiten bestehen)? Gibt es einen Zusammenhang zwischen der Auswahl der Aufgaben und bestimmten Merkmalen der Prüflinge?)
  - Die Ergebnisse hierzu finden sich in den Abschnitten 2.1 (Deutsch), 3.1 (Englisch) und 4.1 (Französisch).
- 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 den Abschnitten 2.2 (Deutsch), 3.2 (Englisch) und 4.2 (Französisch).
- 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 den Abschnitten 2.3 (Deutsch), 3.3 (Englisch) und 4.3 (Französisch).

Stichprobe und Datenerhebung

An der Evaluation für das Prüfungsjahr 2022 nahmen insgesamt 14 Länder teil. Das für die Datenerhebung vom IQB bereitgestellte elektronische Eingabeinstrument wurde von 12 Ländern genutzt. Die Länder Hamburg und Nordrhein-Westfalen übermittelten Daten aus ihren landeseigenen Erhebungen. Nicht in allen der 14 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. In den drei sprachlichen Fächern 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 sechs 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	3736	1966	323	183
Englisch	3846	2169	464	402
Französisch	1040	11	169	-
Mathematik	-	-	459	259

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 Deutsch

2.1 Auswahl der Poolaufgaben

In der schriftlichen Abiturprüfung im Fach Deutsch können die Prüflinge in allen Ländern aus mindestens drei Wahlaufgaben eine Abiturprüfungsaufgabe auswählen. Das Ergebnis dieser Aufgabenauswahl 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. Zusätzlich wurde in einer Metaanalyse aufgabenübergreifend der Zusammenhang zwischen der Auswahl der Aufgaben und den Vorleistungen der Prüflinge betrachtet. Es wurde also untersucht, ob die Vorleistungen der Prüflinge Einfluss auf die Auswahl von Aufgaben der unterschiedlichen Aufgabenarten haben. Die Ergebnisse dieser Metaanalyse sind in Tabelle 2 dargestellt. Hierfür wurde für jedes Land die Differenz ermittelt zwischen der mittleren Halbjahresleistung derjenigen Prüflinge, die eine Aufgabe aus dem Pool gewählt haben, und der mittleren Halbjahresleistung derjenigen Prüflinge, die sich für eine landeseigene Aufgabe entschieden haben. Die Differenzen (gemessen in Notenpunkten) wurden anschließend über jene Länder aggregiert, die Aufgaben einer bestimmten Aufgabenart aus dem Pool entnommen haben. Die Differenzen sowie die untere und obere Grenze des Vertrauensbereichs[2] sind in der folgenden Tabelle separat für jede Aufgabenart und insgesamt (also aufgabenübergreifend) dargestellt. Eine negative Differenz bedeutet, dass Aufgaben aus dem Pool im Durchschnitt von Prüflingen gewählt wurden, die weniger gute Vorleistungen in der Qualifikationsphase erbracht haben als diejenigen Prüflinge, die eine landeseigene Aufgabe wählten. Eine positive Differenz weist darauf hin, dass Aufgaben aus dem Pool von Prüflingen gewählt wurden, die in der Qualifikationsphase bessere Vorleistungen gezeigt haben als diejenigen Prüflinge, die eine landeseigene Aufgabe wählten. Statistisch signifikante Ergebnisse sind mit einem Stern markiert.

Tabelle 2: Länderübergreifende Ergebnisse zur Auswahl der Aufgaben im Fach Deutsch

Aufgabenart[3]	Anzahl der „Studien“	Mittlere Differenz in Notenpunkten	Vertrauensbereich (95%)
EP	4	-1.02*	[-1.98; -0.05]
EL	2	1.02*	[0.47; 1.56]
IL	16	0.02	[-0.18; 0.23]
AP	5	0.49	[-0.20; 1.18]
MI	3	-0.48	[-3.23; 2.27]
MA	2	0.05	[-3.80; 3.90]
insgesamt	32	-0.01	[-0.28; 0.26]

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

Die Ergebnisse der Metanalyse lassen sich wie folgt zusammenfassen:

- Zwischen den Prüflingen, die eine aus dem Pool stammende Aufgabe gewählt haben, und denjenigen, die eine landeseigene Aufgabe bearbeitet haben, wurden (über alle Aufgabenarten hinweg) keine signifikanten Unterschiede in den Vorleistungen gefunden (Differenz: -0.01 NP, nicht statistisch signifikant).
- Poolaufgaben der Art „Erörterung literarischer Texte“ wurden häufiger von Prüflingen gewählt, die im Vergleich zu anderen Schülerinnen und Schülern signifikant bessere Vorleistungen aufweisen (Differenz: +1.02 NP, statistisch signifikant). Demgegenüber wurden Poolaufgaben der Art „„Erörterung pragmatischer Texte“ häufiger von Schülerinnen und Schülern mit weniger guten Vorleistungen gewählt (Differenz: -1.02 NP, statistisch signifikant). Für die übrigen Aufgabenarten fanden sich keine signifikanten Zusammenhänge mit den Vorleistungen der Prüflinge.

2.2 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 empirischen Schwierigkeit von den landeseigenen Aufgaben unterscheiden. Nachfolgend sind die Ergebnisse der hierzu durchgeführten Metaanalysen dargestellt (Tabelle 3, Tabelle 4, Tabelle 5). Hierfür wurden für jedes Land die Differenzen zwischen der mittleren empirischen Schwierigkeit der Aufgaben aus dem Pool und der mittleren empirischen Schwierigkeit der landeseigenen Aufgaben ermittelt. Diese Differenzen (gemessen in Notenpunkten) wurden dann über alle Länder aggregiert, die an der Evaluation teilgenommen haben. In weiteren Analysen wurden Unterschiede zwischen der empirischen Schwierigkeit der Poolaufgaben und der landeseigenen Aufgaben getrennt nach Anforderungsniveaus und Aufgabenarten betrachtet. Die Ergebnisse dieser Analysen sowie die untere und obere Grenze des Vertrauensbereichs[4] 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 3: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Deutsch

	Anzahl der „Studien“	Mittlere Differenz in Notenpunkten	Vertrauensbereich (95%)
insgesamt	32	0.00	[-0.30; 0.30]

Tabelle 4: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Anforderungsniveau im Fach Deutsch

Anforderungsniveau[5]	Anzahl der „Studien“	Mittlere Differenz in Notenpunkten	Vertrauensbereich (95%)
EN	23	-0.09	[-0.46; 0.27]
GN	9	0.39*	[0.04; 0.73]

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

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

Tabelle 5: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Aufgabenarten im Fach Deutsch

Aufgabenart[6]	Anzahl der „Studien“	Mittlere Differenz in Notenpunkten	Vertrauensbereich (95%)
EP	4	-1.22	[-2.11; -0.34]
EL	2	0.89	[-5.46; 7.23]
IL	16	0.11	[-0.17; 0.38]
AP	5	0.42*	[0.04; 0.80]
MI	3	0.13	[-3.81; 4.06]
MA	2	0.03	[-1.88; 1.94]

Die Ergebnisse der Metanalyse 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: 0.00 NP, nicht statistisch signifikant). Allerdings zeigt eine nach Anforderungsniveau differenzierte Betrachtung, dass die Schülerinnen und Schüler im grundlegenden Niveau bei den Poolaufgaben insgesamt etwas besser abschnitten als die Prüflinge, die eine landeseigene Aufgabe gewählt haben (Differenz: 0.39 NP, statistisch signifikant; Differenz für das erhöhte Niveau: -0.09 NP, nicht statistisch signifikant).
- Für die Aufgaben der Art „Analyse pragmatischer Texte“ fallen die Ergebnisse der Prüflinge für die Poolaufgaben signifikant besser aus als für die zum Vergleich herangezogenen landeseigenen Aufgaben (Differenz: +0.42 NP, statistisch signifikant). 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.

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

2.3 Kriteriale Validität der Aufgaben

Als Indikatoren für die kriteriale Validität[7] 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 über jene Länder aggregiert, die an der Evaluation teilgenommen haben. Die so berechneten Validitätskoeffizienten sowie die untere und obere Grenze des Vertrauensbereichs[8] sind in den beiden folgenden Tabellen getrennt für beide Anforderungsniveaus, separat für alle Aufgabenarten und über alle Aufgaben hinweg dargestellt (Tabelle 6, Tabelle 7). 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 6: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Deutsch

	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben[9]			
EN	27	0.77	[0.73; 0.81]
GN	13	0.75	[0.66; 0.82]
EP	4	0.71	[0.60; 0.79]
EL	2	0.86	[-0.92; 0.99]
IL	20	0.77	[0.72; 0.81]
AP	7	0.81	[0.71; 0.87]
MI	4	0.71	[0.64; 0.78]
MA	3	0.71	[0.22; 0.91]
insgesamt	40	0.77	[0.73; 0.80]
landeseigene Aufgaben			
insgesamt	34	0.77	[0.75; 0.80]

[7] 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.

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

[9] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau, EP = Erörterung pragmatischer Texte, EL = Erörterung literarischer Texte, IL = Interpretation literarischer Texte, AP = Analyse pragmatischer Texte, MI = Materialgestütztes Verfassen informierender Texte, MA = Materialgestütztes Verfassen argumentierender Texte.

Tabelle 7: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Deutsch

	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben[10]			
EN	24	0.77	[0.75; 0.79]
GN	12	0.75	[0.64; 0.83]
EP	4	0.73	[0.67; 0.79]
EL	1	0.70	[0.33; 0.88]
IL	18	0.78	[0.74; 0.81]
AP	6	0.82	[0.72; 0.88]
MI	4	0.68	[0.40; 0.84]
MA	3	0.77	[0.59; 0.88]
insgesamt	36	0.77	[0.74; 0.80]
landeseigene Aufgaben			
insgesamt	31	0.80	[0.78; 0.82]

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

- Der in Bezug auf das Validitätskriterium „Halbjahresnoten“ für die Poolaufgaben berechnete Validitätskoeffizient ist als hoch einzustufen (r = .77) und unterscheidet sich nicht statistisch signifikant von dem Kennwert, der für die landeseigenen Poolaufgaben gefunden wurde (r = .77). Die kleinsten (aber dennoch als „hoch“ einzustufenden) Validitätskoeffizienten wurden für die Poolaufgaben der Aufgabenarten „Erörterung pragmatischer Texte“, „Materialgestütztes Verfassen informierender Texte“ und „Materialgestütztes Verfassen argumentierender Texte“ gefunden (r = .71).
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Poolaufgaben 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 = .80, als hoch einzustufen). Der kleinste Validitätskoeffizient wurde für die Poolaufgaben der Art „Materialgestütztes Verfassen informierender Texte“ ermittelt (r = .68, als hoch einzustufen).

[10] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau, EP = Erörterung pragmatischer Texte, EL = Erörterung literarischer Texte, IL = Interpretation literarischer Texte, AP = Analyse pragmatischer Texte, MI = Materialgestütztes Verfassen informierender Texte, MA = Materialgestütztes Verfassen argumentierender Texte.

2.4 Befragung der Lehrkräfte

In der Tabelle 8 sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Deutsch dargestellt. Diese Analysen wurden für folgende Fragebogenitems durchgeführt:

- Einschätzung des Schwierigkeitsgrads der Aufgaben (von „deutlich zu niedrig“ bis „deutlich zu hoch“)
- „Die Aufgabenstellungen sind klar und verständlich formuliert.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)
- „Für die Bearbeitung der Aufgabe ist thematisches Vorwissen erforderlich.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)
- „Die Aufgabe hat es den Schülerinnen und Schülern meines Kurses ermöglicht, die im Unterricht erworbenen Kenntnisse bzw. Kompetenzen einzubringen.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)

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[11] angegeben.

Tabelle 8: Aufgabenübergreifende Ergebnisse zur Befragung der Lehrkräfte im Fach Deutsch

Aspekt	Anzahl der „Studien“	Mittlere stand. Differenz	Vertrauensbereich (95%)
Schwierigkeitsgrad	28	0.03	[-0.17; 0.22]
Aufgabenstellung klar und verständlich	28	0.05	[-0.16; 0.26]
Thematisches Vorwissen	25	-0.45*	[-0.73; -0.17]
Kenntnisse und Kompetenzen aus dem Unterricht	27	-0.24*	[-0.47; -0.01]

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

Die Ergebnisse der zum Fach Deutsch durchgeführten Lehrkräftebefragung lassen sich wie folgt zusammenfassen:

Aufgabenübergreifend betrachtet wurde der Schwierigkeitsgrad der Poolaufgaben im Fach Deutsch sowohl im Allgemeinen als auch im Hinblick auf die zur Verfügung stehende Arbeitszeit und das jeweilige Anforderungsniveau mehrheitlich als „angemessen“ bewertet. Im Vergleich zu den landeseigenen Aufgaben und zu den Klausuren der Qualifikationsphase wurde der Schwierigkeitsgrad der Poolaufgaben in den meisten Fällen als „ungefähr gleich hoch“ eingeschätzt. Dies zeigt sich auch in den Ergebnissen einer Metaanalyse der Schwierigkeitseinschätzungen der Lehrkräfte, bei der aufgabenübergreifend keine statistisch signifikante Differenz zwischen Poolaufgaben und landeseigenen Aufgaben festgestellt wurde (g = 0.02). Als insgesamt sehr positiv wurden die Erwartungshorizonte und Bewertungshinweise beurteilt. Gleiches gilt auch für die Einschätzungen zur Formulierung der Aufgabenstellungen, die aus Sicht der meisten Lehrkräfte klar und verständlich waren (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: g = 0.05, nicht statistisch signifikant). Bei den Poolaufgaben wurde allerdings verstärkt bemängelt, dass die Aufgabenbearbeitung wenig thematisches Vorwissen erfordere (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: g = -0.45, statistisch signifikant, als „mittel“ einzustufen) und es einige Aufgaben den Schülerinnen und Schülern kaum ermöglichten, die im Unterricht erworbenen Kenntnisse bzw. Kompetenzen einzubringen (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: g = -0.24, statistisch signifikant, als „klein“ einzustufen). Sowohl für die Poolaufgaben als auch für die landeseigenen Aufgaben zeigen die Einschätzungen zu den Materialien ein recht heterogenes Bild. Bei einigen Aufgaben fallen diese Einschätzung überwiegend positiv aus, bei anderen werden die sprachliche und inhaltliche Komplexität der Texte sowie deren Umfang als zu hoch angesehen.

Die anhand der geschlossenen Fragebogenitems vorgenommenen Einschätzungen spiegeln sich auch in den optionalen Freitextantworten wider. So wurde die Angemessenheit des Schwierigkeitsgrades der Aufgaben mehrfach positiv hervorgehoben („Die Aufgaben sind anspruchsvoll, aber machbar.“). Allerdings wurde mehrfach bedauert, dass die Aufgaben wenig Bezüge zum Lehrplan des betreffenden Landes und den jeweiligen Lektürelisten aufwiesen.

3 Englisch

3.1 Auswahl der Aufgaben

In den schriftlichen Abiturprüfungen der Länder im Fach Englisch haben die Prüflinge in der Regel weniger Wahlmöglichkeiten als im Fach Deutsch. Im Kompetenzbereich „Hörverstehen“ ist es aus organisatorischen Gründen nicht möglich, Wahlmöglichkeiten anzubieten, da die Schülerinnen und Schüler die Aufgaben gleichzeitig bearbeiten. Im Kompetenzbereich Sprachmittlung sind nur in wenigen Ländern Wahlaufgaben vorgesehen. Im Kompetenzbereich Schreiben können die Schülerinnen und Schüler hingegen in vielen Ländern zwischen zwei oder mehr Aufgaben bzw. Aufgabenblöcken wählen. Die für die Poolaufgaben festgestellten Auswahlhäufigkeiten fallen dabei je nach Land sehr unterschiedlich aus.

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. Zusätzlich wurden nach Kompetenzbereichen differenzierte Analysen durchgeführt, bei denen folgende Vergleiche gezogen wurden:

- Vergleich der bei den Poolaufgaben zum Hörverstehen erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse
- Vergleich der bei den Poolaufgaben zum Schreiben erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse
- Vergleich der bei den Poolaufgaben zur Sprachmittlung erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse
- Vergleich der bei den Poolaufgaben zum Schreiben erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben des Kompetenzbereiches Schreiben erzielten Ergebnisse

3.2 Schwierigkeit der Aufgaben

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

In den folgenden Tabellen sind die Ergebnisse der Metaanalyse zur empirischen Schwierigkeit der Aufgaben im Fach Englisch dargestellt (Tabelle 9, Tabelle 10 und Tabelle 11). Hierfür wurde für jedes Land die Differenz zwischen der empirischen Schwierigkeit der Poolaufgabe(n) und der mittleren empirischen Schwierigkeit der landeseigenen Aufgaben ermittelt. Da im Fach Englisch die Bewertung auf Aufgabenebene je nach Land entweder anhand von NP oder mittels BE erfolgt, wurden die Differenzen standardisiert und über jene Länder aggregiert, die Aufgaben zu einem bestimmten Kompetenzbereich aus dem Pool entnommen

haben. In weiteren Analysen wurden Unterschiede zwischen der empirischen Schwierigkeit der Poolaufgaben und der landeseigenen Aufgaben getrennt nach den beiden Anforderungsniveaus und den drei Kompetenzbereichen betrachtet. 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). *Hedges' g* sowie die untere und obere Grenze des Vertrauensbereichs[12] sind in den folgenden Tabellen dargestellt. Eine negative Differenz bedeutet, dass Prüflinge bei den Poolaufgaben weniger gute Ergebnisse erzielten als bei den landeseigenen Aufgaben. Eine positive Differenz weist darauf hin, dass Prüflinge bei der Bearbeitung von Poolaufgaben bessere Ergebnisse erzielten. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

Tabelle 9: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Englisch

	Anzahl der „Studien"	Stand. Differenz (*g*)	Vertrauensbereich (95%)
insgesamt	20	0.13	[-0.02; 0.27]

Tabelle 10: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Anforderungsniveau im Fach Englisch

Anforderungsniveau[13]	Anzahl der „Studien"	Stand. Differenz (*g*)	Vertrauensbereich (95%)
EN	13	0.12	[-0.10; 0.35]
GN	7	0.10*	[0.04; 0.16]

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

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

Tabelle 11: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Kompetenzbereichen im Fach Englisch

Kompetenz-bereich[14]	Anzahl der „Studien“	Stand. Differenz (*g*)	Vertrauensbereich (95%)
HV (Pool) vs. S/SM (Land)	6	0.35*	[0.12; 0.58]
S (Pool) vs. S/SM (Land)	7	-0.13	[-0.45; 0.20]
SM (Pool) vs. S/SM (Land)	7	0.17*	[0.06; 0.29]
S (Pool) vs. S (Land)	7	-0.08	[-0.42; 0.25]

Die Ergebnisse der Metaanalysen lassen sich wie folgt zusammenfassen:

- Wie bei den vorherigen Evaluationen für die Prüfungsjahr 2017 und 2019 wurde auch für das Prüfungsjahr 2022 kein signifikanter Unterschied zwischen den bei den Aufgaben aus dem Pool erreichten Ergebnissen und den bei den landeseigenen Aufgaben erreichten Ergebnissen festgestellt (*g* = .13, nicht statistisch signifikant). Eine nach Anforderungsniveau differenzierte Betrachtung zeigt, dass die Schülerinnen und Schüler im grundlegenden Niveau (nicht jedoch im erhöhten Niveau) bei den Poolaufgaben insgesamt etwas besser abschnitten als die Prüflinge, die eine landeseigene Aufgabe gewählt haben. Die festgestellte Differenz ist aber als „klein“ einzustufen (*g* = .10, statistisch signifikant).
- Wie bei der Evaluation für das Prüfungsjahr 2017 wurde auch für das Prüfungsjahr 2022 festgestellt, dass die Prüflinge bei den Aufgaben zum Hörverstehen, die in allen Ländern aus dem Pool stammten, signifikant besser abgeschnitten haben als bei den landeseigenen Aufgaben zum Schreiben und zur Sprachmittlung (*g* = .35, statistisch signifikant, als „klein bis mittel“ einzustufen).
- Für den Kompetenzbereich Sprachmittlung wurde ebenfalls festgestellt, dass die Prüflinge bei den Poolaufgaben signifikant bessere Ergebnisse erzielten, als bei den landeseigenen Aufgaben zum Schreiben und zur Sprachmittlung (*g* = .17, statistisch signifikant); die festgestellte Differenz ist aber als „klein“ einzustufen.
- Für die Poolaufgaben des Kompetenzbereiches Schreiben fanden sich keine statistisch signifikanten Unterschiede zu den landeseigenen Aufgaben (in Bezug auf die landeseigenen Aufgaben zum Schreiben und zur Sprachmittlung: *g* = -.13, in Bezug auf die landeseigenen Aufgaben zum Schreiben: *g* = -.08).

[14] HV (Pool) vs. S/SM (Land) = Vergleich der bei den Poolaufgaben zum Hörverstehen erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse; S (Pool) vs. S/SM (Land) = Vergleich der bei den Poolaufgaben zum Schreiben erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse; SM (Pool) vs. S/SM (Land) = Vergleich der bei den Poolaufgaben zur Sprachmittlung erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse; S (Pool) vs. S (Land) = Vergleich der bei den Poolaufgaben zum Schreiben erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben des Kompetenzbereiches Schreiben erzielten Ergebnisse.

3.3 Kriteriale Validität der Aufgaben

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 über jene Länder aggregiert, die an der Evaluation teilgenommen haben. Die so berechneten Validitätskoeffizienten sowie die untere und obere Grenze des Vertrauensbereichs[15] sind in Tabelle 12 und Tabelle 13 getrennt für beide Anforderungsniveaus, separat für die drei Kompetenzbereiche und über alle Aufgaben 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 12: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Englisch

	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben[16]			
EN	23	0.67	[0.61; 0.72]
GN	17	0.72	[0.63; 0.79]
SM	13	0.71	[0.62; 0.79]
S	17	0.71	[0.62; 0.79]
HV	10	0.61	[0.58; 0.63]
insgesamt	40	0.69	[0.64; 0.73]
landeseigene Aufgaben			
insgesamt	39	0.73	[0.68; 0.77]

[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.

[16] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau, SM = Sprachmittlung, S = Schreiben, HV = Hörverstehen.

Tabelle 13: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Englisch

	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben[17]			
EN	21	0.71	[0.64; 0.77]
GN	15	0.76	[0.67; 0.83]
SM	11	0.74	[0.62; 0.83]
S	17	0.76	[0.67; 0.83]
HV	8	0.65	[0.60; 0.69]
insgesamt	36	0.73	[0.68; 0.78]
landeseigene Aufgaben			
insgesamt	35	0.77	[0.70; 0.82]

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

- Der in Bezug auf das Validitätskriterium „Halbjahresnoten“ für die Poolaufgaben berechnete Validitätskoeffizient ist als hoch einzustufen (r = .69) und unterscheidet sich nicht statistisch signifikant von dem Kennwert, der für die landeseigenen Poolaufgaben gefunden wurde (r = .73). Der kleinste (aber dennoch als hoch einzustufende) Validitätskoeffizient wurde für die Poolaufgaben zum Hörverstehen gefunden (r = .61).
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Poolaufgaben ein aggregierter Validitätskoeffizient von r = .73 (als hoch einzustufen) ermittelt. Dieser Koeffizient unterscheidet sich nicht signifikant von dem Kennwert, der für die landeseigenen Aufgaben bestimmt wurde (r = .77, als hoch einzustufen). Der kleinste Validitätskoeffizient wurde wiederum für die Poolaufgaben zum Hörverstehen ermittelt (r = .65, als hoch einzustufen).

[17] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau, SM = Sprachmittlung, S = Schreiben, HV = Hörverstehen.

3.4 Befragung der Lehrkräfte

In der Tabelle 14 sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Englisch dargestellt. Diese Analysen wurden für folgende Fragebogenitems durchgeführt:

- Einschätzung des Schwierigkeitsgrads der Aufgaben (von „deutlich zu niedrig“ bis „deutlich zu hoch“)
- „Die Aufgabenstellungen sind klar und verständlich formuliert.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)
- „Für die Bearbeitung der Aufgabe ist thematisches Vorwissen erforderlich.“ (von „trifft gar nicht zu“ bis „trifft voll zu“) (nur für Aufgaben zum Kompetenzbereich Schreiben abgefragt)
- „Die Aufgabe hat es den Schülerinnen und Schülern meines Kurses ermöglicht, die im Unterricht erworbenen Kenntnisse bzw. Kompetenzen einzubringen.“ (von „trifft gar nicht zu“ bis „trifft voll zu“) (nur für Aufgaben zum Kompetenzbereich Schreiben abgefragt)

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[18] angegeben.

Tabelle 14: Aufgabenübergreifende Ergebnisse zur Befragung der Lehrkräfte im Fach Englisch

Aspekt	Anzahl der „Studien“	Mittlere stand. Differenz	Vertrauensbereich (95%)
Schwierigkeitsgrad	18	0.20	[-0.01; 0.42]
Aufgabenstellung klar und verständlich	22	0.05	[-0.06; 0.15]
Thematisches Vorwissen	8	0.02	[-0.11; 0.15]
Kenntnisse und Kompetenzen aus dem Unterricht	8	0.03	[-0.24; 0.29]

[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.

Auch im Fach Englisch wurde der Schwierigkeitsgrad der Poolaufgaben und der landeseigenen Aufgaben, von wenigen Ausnahmen abgesehen, in Bezug auf die im Rahmen der Lehrkräftebefragung thematisierten Teilaspekte (d. h. im Hinblick auf die Arbeitszeit und das Anforderungsniveau der Prüfung sowie in Relation zu anderen Prüfungsaufgaben und zu den Klausuren der Qualifikationsphase) überwiegend als angemessen eingeschätzt. Darüber hinaus zeigen die Ergebnisse einer aufgabenübergreifenden Metaanalyse, dass der Schwierigkeitsgrad von Poolaufgaben und landeseigenen Aufgaben sehr ähnlich eingeschätzt wird (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: $g = 0.20$, nicht statistisch signifikant) und es dabei auch zwischen den Kompetenzbereichen keine signifikanten Unterschiede gibt. Die Erwartungshorizonte und Bewertungshinweise wurden, ähnlich wie im Fach Deutsch, insgesamt überaus positiv beurteilt. Zudem gaben die meisten Lehrkräfte an, dass die Aufgabenstellungen klar und verständlich formuliert seien (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: $g = 0.05$, nicht statistisch signifikant). Anders als im Fach Deutsch wurde bei den Poolaufgaben nicht häufiger als bei den landeseigenen Aufgaben angegeben, dass die Aufgabenbearbeitung wenig thematisches Vorwissen erfordere und es einige Aufgaben den Schülerinnen und Schülern kaum ermöglichten, die im Unterricht erworbenen Kenntnisse bzw. Kompetenzen einzubringen (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: $g = 0.03$ bzw. $g = 0.03$, jeweils nicht statistisch signifikant). Die Texte bzw. Materialien der Aufgaben wurden insgesamt positiv beurteilt. Bei den Aufgaben im Kompetenzbereich Schreiben wurde allerdings in einigen Fällen angegeben, dass Zusatzmaterialien (z. B. Diagramme, Karikaturen) nicht funktional gewählt worden seien.

Auch die Freitextantworten der Lehrkräfte lassen darauf schließen, dass der Schwierigkeitsgrad der Poolaufgaben als angemessen eingeschätzt wird. So wurde mehrfach hervorgehoben, dass die Poolaufgaben „fair“ und „gut lösbar“ seien. Auch die Klarheit und Verständlichkeit der Aufgabenstellungen wurde in den Freitextantworten nochmals unterstrichen. Vor allem bei den Schreibaufgaben wurde die Auswahl der Texte positiv hervorgehoben; gleichzeitig wurde mehrfach angeregt, „auch andere Textarten wie Comics, Bilder, Diagramme oder Karikaturen zu verwenden“. Bei den Poolaufgaben zum Hörverstehen und zur Sprachmittlung gaben hingegen mehrere kritische Stimmen an, dass die verwendeten Texte nicht schülernah seien. In mehreren Ländern äußerten Lehrkräfte zudem den Eindruck, dass die Aufgaben zum grundlegenden Niveau (in Relation zu den Aufgaben des erhöhten Niveaus) zu anspruchsvoll seien. Außerdem wurde angegeben, dass die auf literarischen Textvorlagen basierenden Schreibaufgaben aus ihrer Sicht anspruchsvoller seien als die Schreibaufgaben zu nicht-literarischen Texten.

4 Französisch

4.1 Auswahl der Aufgaben

In der Regel sind die Abiturprüfungen im Fach Französisch wie im Fach Englisch strukturiert; dementsprechend wurde im Rahmen der Auswertungen analog zu Englisch vorgegangen.

Ähnlich wie im Fach Englisch liegen auch im Fach Französisch zur Auswahl der aus dem Abituraufgabenpool stammenden Aufgaben insgesamt nur wenige Ergebnisse vor. Diese fallen ebenfalls je nach Land sehr unterschiedlich aus.

4.2 Schwierigkeit der Aufgaben

In den nachfolgenden Tabellen sind die Ergebnisse der Metaanalysen zur empirischen Aufgabenschwierigkeit im Fach Französisch dargestellt. Hierfür wurde für jedes Land die Differenz zwischen der empirischen Schwierigkeit der Poolaufgabe(n) und der mittleren empirischen Schwierigkeit der landeseigenen Aufgaben ermittelt. Da im Fach Französisch die Bewertung auf Aufgabenebene je nach Land entweder anhand von NP oder mittels BE erfolgt, wurden die Differenzen standardisiert und über jene Länder aggregiert, die Aufgaben zu einem bestimmten Kompetenzbereich aus dem Pool entnommen haben. In weiteren Analysen wurden Unterschiede zwischen der empirischen Schwierigkeit der Poolaufgaben und der landeseigenen Aufgaben getrennt nach den beiden Anforderungsniveaus und den drei Kompetenzbereichen betrachtet. 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). *Hedges' g* sowie die untere und obere Grenze des Vertrauensbereichs[19] sind in den folgenden Tabellen dargestellt. Eine negative Differenz bedeutet, dass Prüflinge bei den Poolaufgaben weniger gute Ergebnisse erzielten als bei den landeseigenen Aufgaben. Eine positive Differenz weist darauf hin, dass Prüflinge bei der Bearbeitung von Poolaufgaben bessere Ergebnisse erzielten. Statistisch signifikante Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben sind mit einem Stern markiert.

[19] 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 15: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit im Fach Französisch

	Anzahl der „Studien“	Stand. Differenz (*g*)	Vertrauensbereich (95%)
insgesamt	14	0.02	[-0.19; 0.23]

Tabelle 16: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Anforderungsniveau im Fach Französisch

Anforderungsniveau[20]	Anzahl der „Studien“	Stand. Differenz (*g*)	Vertrauensbereich (95%)
EN	12	0.01	[-0.25; 0.27]
GN	2	0.12	[-0.11; 0.36]

Tabelle 17: Aufgabenübergreifende Ergebnisse zur empirischen Schwierigkeit getrennt nach Kompetenzbereichen im Fach Französisch

Kompetenzbereich[21]	Anzahl der „Studien“	Stand. Differenz (*g*)	Vertrauensbereich (95%)
HV (Pool) vs. S/SM (Land)	4	-0.08	[-0.55; 0.40]
S (Pool) vs. S/SM (Land)	4	-0.30	[-1.91; 1.30]
SM (Pool) vs. S/SM (Land)	6	0.20*	[0.04; 0.35]
S (Pool) vs. S (Land)	4	-0.18	[-1.70; 1.35]

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

[21] HV (Pool) vs. S/SM (Land) = Vergleich der bei den Poolaufgaben zum Hörverstehen erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse; S (Pool) vs. S/SM (Land) = Vergleich der bei den Poolaufgaben zum Schreiben erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse; SM (Pool) vs. S/SM (Land) = Vergleich der bei den Poolaufgaben zur Sprachmittlung erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben der Kompetenzbereiche Schreiben und Sprachmittlung erzielten Ergebnisse; S (Pool) vs. S (Land) = Vergleich der bei den Poolaufgaben zum Schreiben erzielten Ergebnisse mit dem Mittelwert der im jeweiligen Land bei den landeseigenen Aufgaben des Kompetenzbereiches Schreiben erzielten Ergebnisse.

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

- Wie bei der vorherigen Evaluation für das Prüfungsjahr 2019 wurde auch für das Prüfungsjahr 2022 kein signifikanter Unterschied zwischen den bei den Poolaufgaben erreichten Ergebnissen und den bei den landeseigenen Aufgaben erreichten Ergebnissen festgestellt (g = .02, nicht statistisch signifikant). Bei einer nach Anforderungsniveaus differenzierten Betrachtung zeigen sich ebenfalls keine statistisch signifikanten Unterschiede.
- Bei einer nach Kompetenzbereichen differenzierten Betrachtung zeigten sich für die Poolaufgaben zum Hörverstehen und Schreiben keine signifikanten Unterschiede zu den jeweils zum Vergleich herangezogenen landeseigenen Aufgaben. Für den Kompetenzbereich Sprachmittlung wurde hingegen festgestellt, dass die Prüflinge bei den Poolaufgaben signifikant besser abschnitten, als bei den landeseigenen Aufgaben zum Schreiben und zur Sprachmittlung (g = .20, statistisch signifikant); die festgestellte Differenz ist aber als „klein“ einzustufen.

4.3 Kriteriale Validität der Aufgaben

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 über jene Länder aggregiert, die an der Evaluation teilgenommen haben. Die so berechneten Validitätskoeffizienten sowie die untere und obere Grenze des Vertrauensbereichs[22] sind in Tabelle 18 und Tabelle 19 getrennt für beide Anforderungsniveaus, separat für die drei Kompetenzbereiche und über alle Aufgaben 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 18: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Halbjahresleistungen in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Französisch

	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben[23]			
EN	17	0.76	[0.68; 0.82]
GN	2	0.80	[0.50; 0.93]
SM	8	0.77	[0.72; 0.82]
S	7	0.82	[0.68; 0.90]
HV	4	0.58	[0.35; 0.74]
insgesamt	19	0.76	[0.69; 0.82]
landeseigene Aufgaben			
insgesamt	25	0.78	[0.69; 0.84]

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

[23] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau, SM = Sprachmittlung, S = Schreiben, HV = Hörverstehen.

Tabelle 19: Aufgabenübergreifende Ergebnisse zur Validität in Bezug auf die Klausurnoten in der Qualifikationsphase getrennt für Poolaufgaben und landeseigene Aufgaben im Fach Französisch

	Anzahl der „Studien“	Validitätskoeffizient	Vertrauensbereich (95%)
Poolaufgaben[24]			
EN	15	0.82	[0.70; 0.89]
GN	-	-	-
SM	6	0.81	[0.70; 0.88]
S	7	0.88	[0.61; 0.97]
HV	2	0.60	[-0.89; 0.99]
insgesamt	15	0.82	[0.70; 0.89]
landeseigene Aufgaben			
insgesamt	21	0.78	[0.70; 0.83]

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

- Der in Bezug auf das Validitätskriterium „Halbjahresnoten“ für die Poolaufgaben berechnete Validitätskoeffizient ist als hoch einzustufen (r = .76) und unterscheidet sich nicht statistisch signifikant von dem Kennwert, der für die landeseigenen Poolaufgaben gefunden wurde (r = .78). Wie im Fach Englisch wurde der kleinste (hier als mittel einzustufende) Validitätskoeffizient für die Poolaufgaben zum Hörverstehen gefunden (r = .58).
- In Bezug auf das Validitätskriterium „Klausurnoten“ wurde metaanalytisch für die Poolaufgaben ein aggregierter Validitätskoeffizient von r = .82 (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 Poolaufgaben zum Hörverstehen ermittelt (r = .60, als mittel einzustufen).

[24] EN = Erhöhtes Niveau, GN = Grundlegendes Niveau, SM = Sprachmittlung, S = Schreiben, HV = Hörverstehen.

4.4 Befragung der Lehrkräfte

In der Tabelle 20 sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Französisch dargestellt. Diese Analysen wurden für folgende Fragebogenitems durchgeführt:

- Einschätzung des Schwierigkeitsgrads der Aufgaben (von „deutlich zu niedrig“ bis „deutlich zu hoch“)
- „Die Aufgabenstellungen sind klar und verständlich formuliert.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)
- „Für die Bearbeitung der Aufgabe ist thematisches Vorwissen erforderlich.“ (von „trifft gar nicht zu“ bis „trifft voll zu“) (nur für Aufgaben zum Kompetenzbereich Schreiben abgefragt)
- „Die Aufgabe hat es den Schülerinnen und Schülern meines Kurses ermöglicht, die im Unterricht erworbenen Kenntnisse bzw. Kompetenzen einzubringen.“ (von „trifft gar nicht zu“ bis „trifft voll zu“) (nur für Aufgaben zum Kompetenzbereich Schreiben abgefragt)

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[25] angegeben.

Tabelle 20: Aufgabenübergreifende Ergebnisse zur Befragung der Lehrkräfte im Fach Französisch

Aspekt	Anzahl der „Studien“	Mittlere stand. Differenz	Vertrauensbereich (95%)
Schwierigkeitsgrad	8	0.33	[-0.10; 0.76]
Aufgabenstellung klar und verständlich	11	0.00	[-0.13; 0.13]
Thematisches Vorwissen	4	-0.50	[-2.01; 1.01]
Kenntnisse und Kompetenzen aus dem Unterricht	4	-0.34	[-1.29; 0.60]

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

Auch im Fach Französisch wurde der Schwierigkeitsgrad der Poolaufgaben und der landeseigenen Aufgaben in Bezug auf die im Rahmen der Lehrkräftebefragung thematisierten Teilaspekte (d. h. im Hinblick auf die Arbeitszeit und das Anforderungsniveau der Prüfung sowie in Relation zu anderen Prüfungsaufgaben und zu den Klausuren der Qualifikationsphase) überwiegend als angemessen eingeschätzt. Der Schwierigkeitsgrad der Poolaufgaben wurde dabei tendenziell als etwas höher eingeschätzt, wobei die im Rahmen einer Metaanalyse ermittelte Differenz zu den landeseigenen Aufgaben (g = 0.33, als „klein bis mittel“ einzustufen) – auch wegen der im Vergleich zu den anderen Fächern geringeren Stichprobe – statistisch nicht signifikant ist. Eine nach Kompetenzbereichen differenzierte Betrachtung ergibt allerdings vorsichtige Hinweise darauf, dass die Poolaufgaben zum Hörverstehen (im Vergleich zu den landeseigenen Aufgaben für die Kompetenzbereiche Schreiben und Sprachmittlung) als schwieriger beurteilt wurden.[26] Darüber hinaus folgen die Ergebnisse der Lehrkräftebefragung für das Fach Französisch einem ähnlichen Muster wie im Fach Englisch. Positiv beurteilt wurden die Erwartungshorizonte und Bewertungshinweise, die Klarheit und Verständlichkeit der Aufgabenstellungen (Differenz zwischen Poolaufgaben und landeseigenen Aufgaben: g = 0.00) sowie die Materialien. Im Hinblick auf die Bezüge zu thematischem Vorwissen (g = -0.50) und zu den im Unterricht erworbenen Kenntnissen bzw. Kompetenzen (g = -0.35) finden sich aufgabenübergreifend zwar jeweils als „klein bis mittel“ einzustufende Differenzen zwischen Poolaufgaben und landeseigenen Aufgaben. Diese Differenzen sind jedoch ebenfalls nicht statistisch signifikant.

In den Freitextfeldern wurde mehrfach der als angemessen betrachtete Schwierigkeitsgrad der Poolaufgaben als positiver Aspekt hervorgehoben. Einige Lehrkräfte bemängelten, dass die Passung zwischen den Aufgaben und den inhaltlichen Vorgaben des jeweiligen landesspezifischen Lehrplans zu gering sei.

[26] Hierzu liegen allerdings nur Daten aus zwei Ländern (Land 3 und Land 4) vor.

5 Mathematik – Befragung der Lehrkräfte

In der Tabelle 21 sind die Ergebnisse von Metaanalysen für die Einschätzungen der Lehrkräfte zu den Poolaufgaben und den landeseigenen Aufgaben im Fach Mathematik dargestellt. Diese Analysen wurden für folgende Fragebogenitems durchgeführt:

- Einschätzung des Schwierigkeitsgrads der Aufgaben (von „deutlich zu niedrig“ bis „deutlich zu hoch“)
- „Die Arbeitsaufträge sind sprachlich eindeutig und verständlich.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)
- „Der Umfang der Aufgabe ist für den zeitlichen Rahmen angemessen.“ (von „trifft gar nicht zu“ bis „trifft voll zu“)

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[27] angegeben.

Tabelle 21: Aufgabenübergreifende Ergebnisse zur Befragung der Lehrkräfte im Fach Mathematik

Aspekt	Anzahl der „Studien“	Mittlere stand. Differenz	Vertrauensbereich (95%)
Schwierigkeitsgrad	69	0.23*	[0.10; 0.36]
Aufgabenstellungen	61	0.06	[-0.07; 0.19]
Umfang der Aufgaben	68	-0.03	[-0.13; 0.08]

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

Im Fach Mathematik ergeben die Ergebnisse der Lehrkräftebefragung für alle drei abgefragten Aspekte ein heterogenes Bild. Der Schwierigkeitsgrad der Poolaufgaben wurde in den meisten Ländern als etwas zu hoch eingeschätzt. Es finden sich jedoch auch Länder, in denen der Schwierigkeitsgrad der Poolaufgaben mehrheitlich als angemessen beurteilt wurde. Auch der Umfang der Aufgaben wurde je nach Land unterschiedlich beurteilt, was vermutlich auch darauf zurückzuführen ist, dass den Prüflingen vor dem Hintergrund der Corona-Pandemie in einigen Ländern zusätzliche Auswahl- bzw. Arbeitszeit gewährt wurde, in anderen hingegen nicht. In den Ländern, in denen der Umfang der Aufgaben kritisch beurteilt wurde, wurde in den Freitextfeldern wiederholt geäußert, dass die „Anzahl der Teilaufgaben zu hoch“ sei und den Prüflingen „wenig Zeit zum Nachdenken“ bliebe. Demgegenüber finden sich jedoch auch Länder, in denen die Lehrkräfte den Umfang der Poolaufgaben überwiegend als angemessen einschätzten. Länderunterschiede zeigen sich auch im Hinblick auf die Einschätzungen zur Formulierung der Aufgaben. In der Mehrheit der Länder beurteilten die Lehrkräfte die Arbeitsaufträge als sprachlich eindeutig und gut verständlich formuliert. Negativ beurteilt wurde dieser Aspekt vor allem in zwei Ländern, wobei in den Freitextantworten wiederholt die Länge der Aufgabenstellung als problematisch angegeben wurde.

Analog zu den Evaluationsergebnissen für das Prüfungsjahr 2021 ist auch für das Prüfungsjahr 2022 festzuhalten, dass Poolaufgaben und landeseigene Aufgaben im Hinblick auf die drei abgefragten Aspekte in den meisten Ländern sehr ähnlich eingeschätzt wurden. Aufgabenübergreifend wurde im Rahmen von Metaanalysen nur in Bezug auf den Schwierigkeitsgrad ein Unterschied zwischen Poolaufgaben und landeseigenen Aufgaben ermittelt. Die Poolaufgaben wurden dabei signifikant schwieriger eingeschätzt; die festgestellte Differenz ist allerdings als „klein“ einzustufen (g = 0.23). In Bezug auf die Einschätzungen zu den Aufgabenstellungen und zum Umfang der Aufgaben findet sich ein solcher Unterschied zwischen Poolaufgaben und landeseigenen Aufgaben nicht (g = -0.03 bzw. g = 0.06).

6 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.