Prognosegüte bildungsstandardbasierter Tests

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

am Fachbereich Erziehungswissenschaft und Psychologie der
Freien Universität Berlin

vorgelegt von

Master of Science
Fuchs, Gesine

Potsdam, Dezember 2018

Erstgutachter:
Prof. Dr. Martin Brunner, Universität Potsdam

Zweitgutachter:
Prof. Dr. Stefan Krumm, Freie Universität Berlin

Datum der Einreichung: 13. Dezember 2018
Tag der Disputation: 25. Januar 2019

Inhalt

Zusammenfassung

Seit 2003 wurde den Ländern in Form der Bildungsstandards für bestimmte Fächer (beispielsweise Mathematik, Deutsch) ein bundesweit einheitlicher Referenzrahmen zur Verfügung gestellt um die Leistungen ihrer Schülerinnen und Schüler einzuordnen (KMK, 2016). Insbesondere bildungsstandardbasierte Tests – in Form der Vergleichsarbeiten (VERA) – sind an Schulen weit verbreitet. So sind Lehrkräfte der 3. und 8. Jahrgangsstufe öffentlicher Schulen in Deutschland dazu verpflichtet, jährlich in mindestens einem Fach die VERA-Tests für ihre insgesamt etwa 1.4 Millionen Schülerinnen und Schüler durchzuführen (KMK, 2012 Fassung von 2018; Statistisches Bundesamt, 2017). Auf der Grundlage der VERA-Tests erhalten Lehrkräfte Leistungsinformationen auf der Schul-, Klassen- und Individualebene, welche primär für die Unterrichts- und Schulentwicklung genutzt werden sollen (KMK, 2015). Die Bildungsstandards sollen „ fachliche und fachübergreifende Basisqualifikationen [formulieren], die für die weitere schulische und berufliche Ausbildung von Bedeutung sind und die anschlussfähiges Lernen ermöglichen.“ (KMK, 2004, S. 7). Dieser Anspruch impliziert, dass auf Grundlage der Testergebnisse zentrale Kriterien des Schulerfolgs vorhersagbar sein sollten. Es existierte lediglich eine Studie (Graf, Harych, Wendt, Emmrich & Brunner, 2016), die die Prognosegüte bildungsstandardbasierter Tests explizit untersucht hat.

Infolge dessen wurde die übergreifende Forschungsfrage untersucht, inwiefern bildungsstandardbasierte Mathematik- und Deutschtests (im Lesen) geeignet sind, den zukünftigen schulischen Erfolg von Schülerinnen und Schüler – fokussiert auf spätere Schulhalbjahresnoten und bildungsstandardbasierte Testleistungen im selben Fach – vorherzusagen.

Die zentralen Befunde der vorliegenden Dissertation lassen sich wie folgt zusammenfassen:

(1) Die Prognosegüte auf bis zu 5 Jahre spätere Testleistungen und Noten ist vergleichbar mit jener für kommerziell erhältliche Schulleistungstests.
(2) Bildungsstandardbasierte Tests haben einen inkrementellen prognostischen Mehrwert gegenüber Schulhalbjahresnoten auf bis zu 5 Jahre spätere Testleistungen und Noten.
(3) Die Ausprägung der Prognosegüte bildungsstandardbasierter Tests und deren inkrementeller prognostischer Mehrwert variiert zum Teil zwischen den Schulen. An den meisten Schulen liefern die bildungsstandardbasierten Tests jedoch einen Informationsgewinn zur Prognose.

(4) Bildungsstandardbasierte Tests können im Sinne eines Screenings eingesetzt werden, um die Identifikation von Schülerinnen und Schüler zu verbessern, die bis zu 5 Jahre spätere Bildungsergebnisse in Form von bildungsstandardbasierten Testleistungen und Noten verfehlen.

Damit wird ein Beitrag zu bestehenden Forschungsdesideraten in Deutschland geleistet, die nicht nur für bildungsstandardbasierte Tests bestehen sondern ebenso für standardisierte Schulleistungstests im Allgemeinen. Darüber hinaus werden verschiedene Ansätze diskutiert, anhand derer zukünftig die Validierung und Implementierung von VERA-Tests im Sinne einer formativen Leistungsmessung an Schulen gefördert werden könnte.

Summary

Since 2003 a national educational standard was introduced for specific subjects (e. g. mathematics, german) in primary and secondary education in Germany to evaluate the performance of their students (KMK, 2016). Standard-based tests that are used in the state-wide assessment VERA are very common in German schools: Teachers of Grades 3 and 8 at public schools have to test their students in at least one subject every year. Consequently, about 1.4 million students complete the VERA assessment every year in Germany (KMK, 2012 Version of 2018; Statistisches Bundesamt, 2017). The state-wide assessment program VERA offers teachers information on students' proficiency levels on the school-, class- and student-level. Teachers are supposed to use this information primarily to improve their teaching and for school improvement (KMK, 2015). The national educational standard should describe basic competencies that are important for students' future success at school and the working world as well as for their future learning (KMK, 2004). This implies that their current proficiency levels are presumed to inform about their future school success. So far existed one study (Graf et al., 2016) that focused on the prediction validity of national educational standard assessment explicitly.

In this light, the present PhD thesis examined the extent to which standard-based test scores in mathematics and German reading comprehension predict students' success at school as expressed in their future grades and standard-based proficiency levels in the corresponding subject.

The major contributions of this PhD project are summarized as follows:

(1) The power of standard-based test scores to predict future standard-based test scores and grades up to 5 years later is comparable to the predictive power of commercial achievement tests in Germany.
(2) Standard-based test scores incrementally predict future standard-based test scores and grades up to 5 years later beyond controlling for grades.
(3) The predictive power of standard-based test scores vary to some extent between schools, but standard-based tests are predictive of students' future test scores and grades at most schools.
(4) Standard-based tests can be used as a screening to improve the identification of students who are at risk of failing important educational outcomes as assessed in proficiency levels and grades up to 5 years later.

This kind of evidence contributes to reduce the existing lack in the current state of research for standard-based tests in Germany that is not limited to the national educational standard. Furthermore, approaches to enhance the validity and implementation of standard-based tests in state-wide assessment programs like VERA as formative learning assessments are discussed.

1 Relevanz standardisierter Leistungsmessung an Schulen

Im Zuge des schlechten Abschneidens der Schülerinnen und Schüler in Deutschland in umfassenden internationalen Schulleistungsstudien, wie z. B. TIMSS (Trends in International Mathematics and Science Study) und PISA (Programme for International Student Assessment), wurden zahlreiche Maßnahmen eingeleitet, um die Qualität des deutschen Schulwesens zu verbessern. Ein wesentlicher Schritt war in diesem Zusammenhang die Verabschiedung der länderübergreifenden Bildungsstandards durch die Kultusministerkonferenz (KMK) in den Jahren 2003 und 2004 für bestimmte Klassenstufen im Primar- und Sekundarbereich (Hauptschulabschluss und Mittlerer Schulabschluss). 2012 wurden diese um Bildungsstandards für die Allgemeine Hochschulreife ergänzt (KMK, n. d.). Damit einhergehend hat die KMK im Jahre 2006 eine Gesamtstrategie zum Bildungsmonitoring beschlossen. Im Rahmen dieser wird das Ziel verfolgt, die systematische Umsetzung der Maßnahmen zur Qualitätssicherung auf Schulebene bis auf Ebene des gesamten Bildungssystems zu unterstützen, um so u. a. zur Sicherung und Entwicklung der Bildungsqualität beizutragen (KMK, 2015). So sind seit Beginn des 21. Jahrhunderts Lehrkräfte in zunehmenden Maße aufgrund von externen Vorgaben mit verschiedenen Formen der Leistungsfeststellung konfrontiert (Pant, Emmrich, Harych & Kuhl, 2011). Im Vergleich zu anderen OECD-Ländern nahmen Schülerinnen und Schüler in Deutschland im Jahre 2015 jedoch am seltensten an verpflichtenden standardisierten Leistungstests teil (Organisation for Economic Co-operation and Development [OECD], 2016). Ebenfalls wird im Vergleich mit anderen OECD-Ländern deutlich, dass in Deutschland freiwillige standardisierte Leistungstests seltener als verpflichtende Leistungstests von Lehrkräften an Schulen eingesetzt werden (OECD, 2016).

Zusammenfassend kann festgehalten werden, dass national und international ein datenbasierter Ansatz verfolgt wird, mit dem Ziel, die Leistung der Schülerinnen und Schüler zu verbessern (Ikemoto & Marsh, 2007; Mandinach, 2012). Damit ist die Hoffnung verbunden, dass gezieltere Entscheidungen in Politik und Praxis getroffen werden können, um letztendlich das Lernen von Schülerinnen und Schülern an Schulen zu fördern und so eine Leistungssteigerung auf Seiten der Schülerinnen und Schüler zu erzielen.

Als eines der wirksamsten Rahmenkonzepte zur Förderung schulischen Lernens gilt in diesem Zusammenhang die formative Leistungsmessung (Schütze, Souvignier & Hasselhorn, 2018).

Die formative Leistungsmessung „[...] bezeichnet die lernbegleitende Beurteilung von Schülerleistung mit dem Ziel, diagnostische Informationen zu nutzen, um Unterricht und Lernen zu verbessern.“ (Schütze et al., 2018, S. 697). Insbesondere betonen Black und Wiliam (2009) die entscheidungsunterstützende Funktion formativer Leistungsmessung, indem die gewonnenen Leistungsinformationen die Grundlage für Entscheidungen in Bezug auf nachfolgende Lehr- und Lernprozesse bilden.

Ansätze zur datenbasierten Verbesserung von Schulqualität und Schülerleistungen beschränken sich zwar nicht nur auf die Leistungsmessung allein (für einen Überblick s. Wurster et al., 2017). Im Vergleich zu anderen Maßnahmen zur Datengewinnung (z. B. interne Evaluation) steht die standardisierte Leistungsmessung stärker im (Forschungs-) Fokus, da Schulen zu einer regelmäßigen Durchführung, wie beispielsweise in Form bildungsstandardbasierter Tests wie VERA, verpflichtet werden.

1.1 Kriteriumsorientierte Leistungsmessung mit bildungsstandardbasierten Tests

Bildungsstandardbasierte Testverfahren, also Tests, bei deren Testentwicklung die Bildungsstandards den theoretischen Rahmen vorgeben, sind in Deutschland weit verbreitet (Pant, Tiffin-Richards & Stanat, 2017). Dies liegt darin begründet, dass im Rahmen der bereits erwähnten Gesamtstrategie zum Bildungsmonitoring standardisierte Leistungstests zur Messung der Bildungsstandards festgelegt wurden. In Deutschland werden sowohl nationale Ländervergleichstests als auch bundesweite Vergleichsarbeiten/Lernstandserhebungen (VERA; KMK, 2015) zur Messung der Bildungsstandards durchgeführt. Insbesondere die VERA-Tests sind an deutschen Schulen weit verbreitet. So sind Lehrkräfte der 3. und 8. Jahrgangsstufe öffentlicher Schulen in Deutschland dazu verpflichtet, jährlich in mindestens einem Fach bildungsstandardbasierte Tests – in Form der Vergleichsarbeiten (VERA) – für ihre insgesamt etwa 1.4 Millionen Schülerinnen und Schüler durchzuführen (KMK, 2012 Fassung von 2018; Statistisches Bundesamt, 2017). Außerhalb dieser verpflichtenden Testungen in allgemein festgelegten Testzeiträumen werden Lehrkräften bildungsstandardbasierte Testaufgaben aus vergangenen VERA-Testungen z. T. über Aufgabendatenbanken zur Verfügung gestellt (bspw. Institut für Schulqualität der Länder Berlin und Brandenburg e.V. [ISQ], n. d.). Neben diesen unentgeltlichen Angeboten gibt es zudem bildungsstandardbasierte Testhefte, die Lehrkräfte kommerziell erwerben können.

Das besondere Potenzial bildungsstandardbasierter Tests zur Leistungsmessung wird in der kriterialen Bezugsnormorientierung dieser Tests gesehen (Groß Ophoff, 2013; Isaac, 2013). Tests zur kriteriumsorientierten Leistungsmessung stellen innerhalb der pädagogisch-psychologischen Diagnostik seit den 1970er-Jahren ein zentrales Forschungsgebiet dar (Ingenkamp & Lissmann, 2008; Klauer, 1987). So ist es mit bildungsstandardbasierten Tests möglich, die Leistungen von Schülerinnen und Schüler auf dem kriterialen Maßstab der Bildungsstandards zu verorten. Auf diese Weise könnten Lehrkräfte feststellen, in welchem Ausmaß die festgelegten bundesweiten Leistungserwartungen von ihren Schülerinnen und Schülern zu definierten Zeitpunkten in der Bildungskarriere erreicht werden. Diese Leistungsinformationen können kriterial in Bezug auf die Bildungsstandardmetrik durch die Einordnung (a) des erreichten Punktwertes auf der kontinuierlichen Kompetenzskala (Mittelwert für Gesamtdeutschland $M = 500$, $SD = 100$) und (b) der erreichten Kompetenzstufe (i. d. R. in *Unter Mindeststandard*, *Mindeststandard*, *Regelstandard*, *Über Regelstandard* und *Optimalstandard*) auf der Grundlage der Kompetenzstufenmodelle interpretiert werden.

1.2 Vergleichsarbeiten (VERA)

Bildungsstandardbasierte Leistungstests in Form der bundesweiten Vergleichsarbeiten/Lernstandserhebungen (VERA) sind die am weitesten verbreiteten und am häufigsten eingesetzten standardisierten Schulleistungstests in Deutschland (Überblick Pant et al., 2017; Wurster et al., 2017). So werden beispielsweise VERA-Tests in der 3. Klassenstufe (VERA 3) und 8. Klassenstufe (VERA 8) u. a. in den Fächern Mathematik und Deutsch (für verschiedene Leistungsdomänen wie z. B. Lesen) durchgeführt. Die VERA-3-Tests basieren auf den Bildungsstandards, die für die 4. Klassenstufe der Primarstufe definiert sind und werden somit ein Jahr vor dem definierten Bildungsstandard gemessen. Die VERA-8-Tests basieren auf den Bildungsstandards, die für die 10. Klassenstufe der Sekundarstufe (für den mittleren Schulabschluss) definiert sind und werden somit 2 Jahre vor dem definierten Bildungsstandard gemessen.

Auf der Grundlage der VERA-Tests erhalten Lehrkräfte Leistungsinformationen auf der Schul-, Klassen- und Individualebene, welche primär für die Unterrichts- und Schulentwicklung genutzt werden sollen (KMK, 2015). In Deutschland werden die VERA-Tests eher einem *low-stakes* Kontext zugeordnet (Hellrung & Hartig, 2013; Pant et al., 2017). Das bedeutet, dass Schülerinnen und Schüler oder Lehrkräfte auf der Grundlage der

Leistungsergebnisse nicht mit bindenden Konsequenzen, wie z.B. Sanktionen, zu rechnen haben. Im Vergleich zu anderen Verfahren zur datenbasierten Schul- und Unterrichtsentwicklung, wie beispielsweise der Schulinspektion, internen Evaluation oder dem Mittleren Schulabschluss, schätzen Schul- und Fachkonferenzleitungen die Nützlichkeit und Diagnosegüte von VERA am niedrigsten ein (Wurster et al., 2017). Insbesondere Lehrkräfte bewerten die Nützlichkeit von VERA im Vergleich zu Schul- und Fachkonferenzleitungen am niedrigsten (Wurster et al., 2017).

Auch wenn das Potenzial bildungsstandardbasierter Tests insbesondere in der kriterialen Bezugsnorm gesehen wird, wird die Bildungsstandardmetrik in den meisten Bundesländern nicht für die Ergebnisrückmeldung im Rahmen von VERA an Lehrkräfte genutzt (Isaac, 2013). Zudem erhalten Lehrkräfte nur in einigen Bundesländern eine kriteriale Einordnung der individuellen Leistungen ihrer Schülerinnen und Schüler (Pant et al., 2017). So wird Lehrkräften beispielsweise eine kriteriale Einordnung der individuellen Leistungen ihrer Schülerinnen und Schüler gegeben, indem ihnen der individuelle Punktwert einer Schülerin bzw. eines Schülers auf der Bildungsstandardmetrik unter Einordnung in die Kompetenzstufenbereiche mit einem Unsicherheitsintervall (bspw. 95 %-Intervall) mitgeteilt wird (Pant et al., 2017). Das Unsicherheitsintervall verdeutlicht, den Grad der Ungenauigkeit der Leistungsmessung auf Ebene der einzelnen Schülerinnen und Schüler. Unter Umständen sind die Leistungsmessungen mit großen Messfehlern behaftet, wodurch sich diese Intervalle z.T. über drei Kompetenzstufen erstrecken können. Deshalb wird aus wissenschaftlicher Sicht das Potenzial bildungsstandardbasierter Tests zur individualdiagnostischen Anwendung entweder stark angezweifelt (Pant et al., 2011, 2017) oder vor allem darin gesehen, dass diese eher im Sinne eines Screenings erste Hinweise für eine weiterführende Diagnostik geben können (Köller & Reiss, 2013; Leutner, Fleischer, Spoden & Wirth, 2008).

1.3 Abgrenzung von VERA zu kommerziell erhältlichen standardisierten Schulleistungstests

Neben bildungsstandardbasierten Tests wie VERA stehen Lehrkräften an Schulen auch kommerziell erhältliche standardisierte Testverfahren zur Verfügung, wie beispielsweise der Deutsche Mathematiktest für zweite Klassen (DEMAT 2+; Krajewski, Liehm & Schneider, 2004) oder der Leseverständnistests für Erst- bis Sechstklässler (ELFE 1-6; Lenhard & Schneider, 2006). Ein systematischer Vergleich bildungsstandardbasierter Mathematiktests

mit etablierten standardisierten Schulleistungstests im Fach Mathematik haben Köller und Reiss (2013) vorgenommen. Dabei haben sie festgestellt, dass sich die Tests vor allem in den theoretischen Grundlagen und den diagnostischen Anlässen unterscheiden und weniger in der konkreten Formulierung der Items oder in Kennwerten im Zusammenhang mit der Validität der Tests. Detaillierte empirische Vergleichsstudien liegen für bildungsstandardbasierte Mathematikaufgaben und Mathematikaufgaben aus dem DEMAT 3+ sowie dem DEMAT 4 vor (Winkelmann, Robitzsch, Stanat & Köller, 2012). Dabei konnte festgestellt werden, dass mit Korrelationen von $r = .69$ (DEMAT 3+) und $r = .67$ (DEMAT 4) die bildungsstandardbasierten Mathematikaufgaben und die Aufgaben des standardisierten Schulleistungstests inhaltlich voneinander abgrenzbar sind und nicht exakt dasselbe messen. Ab dem DEMAT 5+, 6+ und 9 wurden laut Aussage der Autoren ebenso die Bildungsstandards der Konzeption zugrunde gelegt (Götz et al., 2013a, 2013; Schmidt et al., 2013). Jedoch liegen bisher keine empirischen Befunde für den Zusammenhang des DEMAT 5+, 6+ und 9 mit bildungsstandardbasierten Tests vor.

Kommerziell erhältliche standardisierte Schulleistungstests werden – im Gegensatz zu bildungsstandardbasierten Tests – insbesondere zur Einzelfalldiagnostik eingesetzt, um beispielsweise den Förderbedarf von Schülerinnen und Schülern zu identifizieren. So sind standardisierte Schulleistungstests vor allem für die Identifizierung von Rechenschwächen bzw. Rechenstörungen geeignet, da sie vor allem im unteren Leistungsbereich differenzieren und Fehleranalysen ermöglichen (Köller & Reiss, 2013). Als weiterer Unterschied kommt hinzu, dass in diesen standardisierten Schulleistungstests die Leistung des Einzelnen im Vergleich zur sozialen Bezugsnorm eingeordnet wird. Diese soziale Bezugsnorm lässt meist eine von zwei Klassifikationen zu: „auffällig“ und „nicht auffällig“. Kinder werden als auffällig eingestuft, wenn sie im Vergleich zu den anderen Kindern die schwächsten Kompetenzen erreichen. Konkret werden Kinder meist als auffällig eingestuft, wenn die Leistung des Kindes den Prozentrang von 15 nicht überschreitet, also lediglich 15 % aller getesteten Kinder eine schwächere oder gleich schwache Leistung aufzeigen (Überblick über den Forschungsstand standardisierter Leistungstests im Elektronischen Supplement [ESM] 1 der Studie III). Kinder, deren Leistung den Prozentrang von 15 überschreitet, werden als „nicht auffällig“ klassifiziert. Potenziell lassen sich die Leistungen von Schülerinnen und Schülern mit bildungsstandardbasierten Tests auf Basis der entwickelten Kompetenzstufenmodelle nicht nur in zwei Klassifikationen, sondern in meist fünf

sogenannte Kompetenzstufen einordnen: (1) Unter Mindeststandard, (2) Mindeststandard, (3) Regelstandard, (4) Über Regelstandard und (5) Optimalstandard.

1.4 Abgrenzung von VERA zu Schulnoten

In der Regel werden an den meisten Schulen in Deutschland die Leistungen von Schülerinnen und Schülern mit Schulnoten auf einer sechsstufigen Skala von 1 = *sehr gut* bis 6 = *ungenügend* bewertet (Kulow, 2011). Aus der bisherigen Forschungsliteratur wird deutlich, dass Schulnoten nur bedingt kriterial vergeben werden und Lehrkräfte die Notengebung an das Leistungsniveau der Schülerschaft anpassen (Hochweber, 2010; Lintorf, 2012; Schmid, Paasch & Katstaller, 2016). Das kann zur Folge haben, dass Schülerinnen und Schüler bei gleicher Leistung in leistungsstarken Lernverbänden schlechtere Noten und in leistungsschwächeren Lernverbänden bessere Noten erhalten. Zudem gibt es zahlreiche weitere empirische Hinweise auf leistungsferne Faktoren, die zu Verzerrungen bei der Notenvergabe führen können, wie beispielsweise der sozioökonomische Status der Eltern (Cicmanec, Johanson & Howley, 2001) oder das Geschlecht der Schülerinnen und Schüler (u. a. Lintorf, 2012; Schreiner, Breit & Haider, 2008). Diese potenziellen Verzerrungen bei der Notenvergabe könnten u. a. als mögliche Erklärung für Unterschiede zwischen Testleistungen und Schulnoten dienen. Hinzu kommt speziell beim Vergleich von Schulnoten mit bildungsstandardbasierten Tests, dass die Tests eine geringere curriculare Nähe besitzen als Schulnoten. Während sich Schulnoten auf die unmittelbar vermittelten Lerninhalte beziehen, messen bildungsstandardbasierte Tests Kompetenzen, die mehr oder weniger mit den Lehrplänen der einzelnen Bundesländer konvergieren und somit eine mehr oder weniger stark ausgeprägte curriculare Nähe aufweisen können (Pant et al., 2011). Folglich könnten Unterschiede zwischen bildungsstandardbasierten Tests und Schulnoten auch auf diesbezügliche Unterschiede zurückgeführt werden.

Trotz dieser Unterschiede geht aus bisherigen Korrelationsstudien hervor, dass bildungsstandardbasierte Testaufgaben und VERA-Tests mit korrespondierenden Schulnoten innerhalb der jeweiligen Fächer Mathematik und Deutsch zwischen $.41 \leq r \leq .71$ miteinander korrelieren (Winkelmann et al., 2012; C. Nachtigall, persönl. Mitteilung der Kennwerte aus den Länderberichten für die aufeinanderfolgenden Kohorten 2009/10 bis 2013/14, 22.10.2018). Die Höhe der Korrelationskennwerte zwischen Schulnoten und bildungsstandardbasierten Tests veranschaulicht, dass diese nur zum Teil ähnliche Leistungsinformationen enthalten. Demnach kann zusammengefasst werden, dass

bildungsstandardbasierte Testaufgaben eine ähnliche, aber nicht identische Schülerleistung im Vergleich zu Schulnoten und standardisierten Schulleistungstests erfassen.

2 Validierung bildungsstandardbasierter Tests

Damit Lehrkräfte an Schulen zielführende, datenbasierte Entscheidungen und Handlungen für das Lernen ihrer Schülerinnen und Schüler treffen können, benötigen sie valide Testdaten. Studien, die solche Annahmen zur Validität von Testverfahren prüfen, haben eine lange Tradition (u. a. Shadish, Cook & Campbell, 2002). So wurden über die Zeit unterschiedlichste Definitionen und Systematisierungen des Begriffs der Validität vorgenommen, deren aktueller Stand in den *Standards for Educational and Psychological Testing* verzeichnet ist (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME]; 2014).

> So wird im Rahmen dieser *Standards* (AERA et al., 2014) die Validität folgendermaßen definiert:
>
> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (S. 13)

Trotz dieser vorliegenden Richtlinien, welche letztlich eine Zusammenfassung von Standards zur Untersuchung verschiedenster Aspekte der Validität darstellen, liegen erst seit wenigen Jahren konkrete Vorschläge für mögliche Rahmenmodelle zur systematischen Untersuchung der Validität von Testverfahren im Schulkontext vor (bspw. Pant et al., 2017; Pellegrino, DiBello & Goldman, 2016). Im Rahmen der vorliegenden Dissertation möchte ich einen argumentationsbasierten Ansatz zur Validierung aufgreifen. Dieser Ansatz stellt im Folgenden unser Arbeitsmodell dar, um eine systematische konzeptionelle und empirische Bewertung der Testwertinterpretation oder -nutzung bildungsstandardbasierter Tests in Bezug auf die Prognosegüte im Rahmen unserer drei Studien vorzunehmen.

Die Grundlage für die nachfolgenden Erläuterungen bieten die Ausführungen von Kane (2013, 2017). Kane orientiert sich in seinen Ausführungen zum argumentbasierten Ansatz an *Toulmin's Model of Inference* (Toulmin, 1958, 2001). Diesen Ansatz habe ich herausgegriffen, da dieser (a) eine Untersuchung der Validität im Sinne der Definition von

Validität im Rahmen der *Standards for Educational and Psychological Testing* ermöglicht, (b) übersichtlich ist und (c) flexibel auf unterschiedlichste Ziele und Kontexte des Einsatzes von Leistungstests adaptiert werden kann. Auch wenn sich unsere Studien ebenso auf der Grundlage der *Standards for Educational and Psychological Testing* begründen lassen, ermöglicht dieser argumentbasierte Ansatz eine systematische Evaluierung unserer Studien zur Validität (konkret mit dem Fokus auf die Prognosegüte) von bildungsstandardbasierten Tests.

Nachfolgend wird der argumentbasierte Ansatz auf der Grundlage der Ausführungen von Kane (2013, 2017) auf der Basis des *Toulmin's Model of Inference* kurz dargestellt und nachfolgend unsere drei Validierungsstudien in dieses Modell eingeordnet.

2.1 Argumentbasierter Ansatz zur Validierung von Leistungstests

Zur Validierung von Leistungstests im Schulkontext nutzt Kane einen argumentbasierten Ansatz, welcher auf *Toulmins Model of Inference* basiert (s. Abbildung 1). Kanes argumentbasierter Ansatz setzt sich aus zwei konzeptionell aufeinanderfolgenden Schritten zusammen (Kane, 2013). Im ersten Schritt muss das sogenannte Argument explizit als Schlussfolgerung, Interpretation, sowie einhergehende Annahme, die auf der Grundlage der Testergebnisse im Leistungstest getroffen werden sollen, formuliert werden (*Interpretation/Use Argument* [IUA]). Das IUA wird als valide angesehen, wenn dieses klar, kohärent und vollständig ist; sowie die darauf basierten Schlussfolgerungen und Annahmen plausibel sind. Im zweiten Schritt kann zwar das IUA nicht bewiesen werden, aber auf der Grundlage relevanter, theoretischer und empirischer Evidenz in Bezug auf seine Klarheit, Kohärenz und Plausibilität evaluiert werden (*Validity Argument*). Nachfolgend werde ich in Anlehnung an die Ausführungen von Kane (2013, 2017) zusammenfassend die einzelnen Komponenten der beiden Schritte knapp erläutern und darauffolgend diese Schritte auf die Validierung der Prognosegüte im Rahmen der drei Studien dieser Dissertationsschrift anwenden. In Abbildung 1 wird *Toulmin's Model of Inference,* mit den jeweiligen einzelnen Komponenten dargestellt.

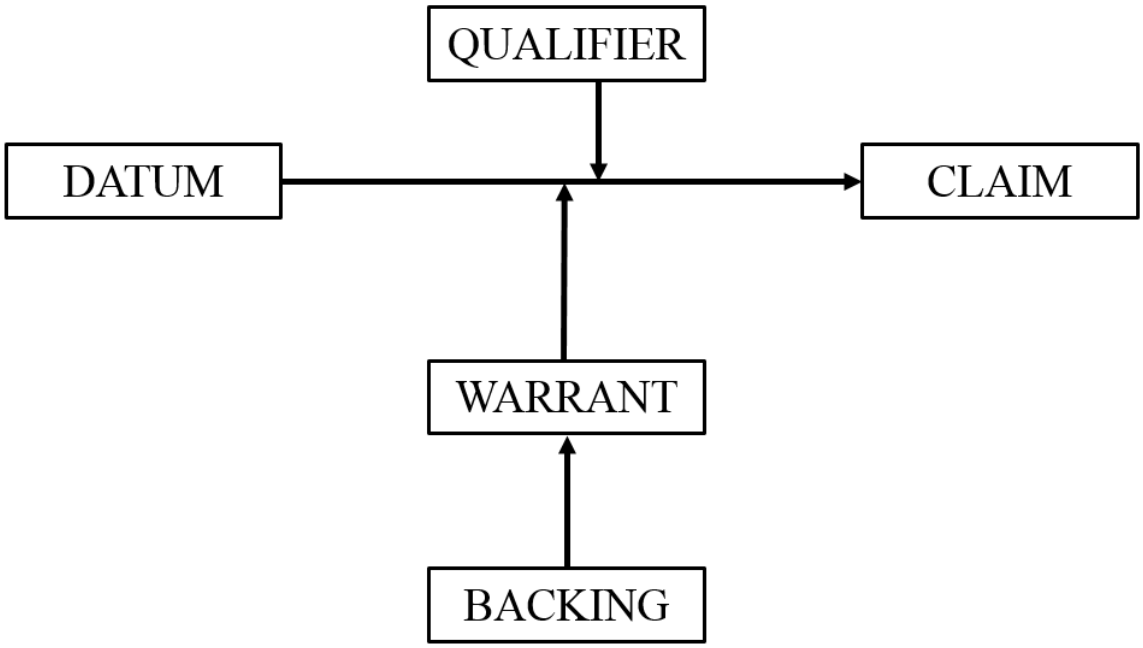


Abbildung 1. Toulmin's Model of Inference (eigene Darstellung mit Genehmigung von Springer Nature Customer Service Centre GmbH: Springer [Standard setting: Bridging the worlds of policy making and research, Pant, Tiffin-Richards & Stanat, Abbildung 4. 2, S. 57] Copyright 2017 bei Springer.

Der Ausgangspunkt jeder Schlussfolgerung stellt das *datum* dar, welches sich auf eine Behauptung bzw. Anspruch (*claim*) bezieht. Das *datum* kann beispielsweise das Ergebnis in einem Mathematikleistungstest sein. Der *claim* könnte darin bestehen, eine Schlussfolgerung in Bezug auf die zukünftige schulische Mathematikleistung treffen zu können. Diese Schlussfolgerungen, werden meist in Form von „wenn ..., dann ..." -Formulierungen getroffen. Hier könnte diese somit konkret lauten: wenn Schülerinnen und Schüler bessere Leistungen im Mathematikleistungstest haben, dann sollten diese zukünftig auch bessere Schulnoten in Mathematik erreichen. Diese Schlussfolgerung beruht auf einer allgemeinen Regel (*warrant*), um von einem *datum* auf ein *claim* zu schließen. Ein *warrant* zur Prognose zukünftiger Leistungen auf der Grundlage eines vorliegenden Testwertes kann beispielsweise die Berechnung von Regressions- oder Korrelationsanalysen darstellen, die letztlich in Form von z. B. Korrelationskoeffizienten die Ausprägung dieses *warrants* wiederspiegeln. Diese *warrants* werden durch empirische Evidenz (*backing*) gestützt. Das Ausmaß der notwendigen empirischen Evidenz zur Rechtfertigung der Schlussfolgerung hängt unter anderem davon ab, wie bedeutungsvoll die Behauptung in Hinblick auf ihre möglichen Konsequenzen ist. So wäre beispielsweise im Rahmen von *high-stake*-Testungen, wie z. B. Klassenarbeiten, einer Behauptung ein deutlich größeres Gewicht beizumessen, wenn in Abhängigkeit des

Testresultates beispielsweise Schülerinnen und Schüler mit ernsthaften Konsequenzen, wie z. B. Belohnungen oder Sanktionen, rechnen müssten. Folglich wäre im Rahmen von *high-stake*-Testungen mehr empirische Evidenz zu fordern als im Kontext von *low-stake*-Testungen. *Warrants* sind nicht unfehlbar. So können andere Faktoren (*qualifier*) vorliegen, die letztlich die Ausprägung des *warrants* beeinflussen können. Somit ist möglicherweise von der Muttersprache des getesteten Kindes abhängig, wie stark der Zusammenhang (z. B. der Korrelationskoeffizient) zwischen der erstmaligen und späteren Leistungstestmessung im Mathematiktest ausgeprägt ist. Ein *claim* kann zudem mehrere (sequentiellen) Schlussfolgerungen einschließen. So können vorherige Schlussfolgerungen in Bezug auf den *claim* als Ausgangspunkt für nachfolgende Schlussfolgerungen dienen.

2.2 Validierungsstudien des Dissertationsprojektes zur Prognosegüte bildungsstandardbasierter Tests

In der vorliegenden Dissertation haben ich im Rahmen dreier Validierungsstudien die übergreifende Forschungsfrage untersucht, inwiefern bildungsstandardbasierte Tests geeignet sind den zukünftigen schulischen Erfolg von Schülerinnen und Schülern – fokussiert auf spätere Schulnoten und bildungsstandardbasierte Testleistungen im selben Fach – vorherzusagen. Die diesbezüglichen Analysen haben wir über die verschiedenen Studien hinweg unter dem Begriff der Prognosegüte (bzw. Prognosekraft, Vorhersagekraft) zusammengefasst. In Abbildung 2 wird dargestellt, wie sich die drei Studien konzeptionell in Bezug auf die verschiedenen Komponenten in *Toulmin's Model of Inference* auf der Grundlage der Erläuterungen von (2013, 2017) einordnen lassen.

In einer der ersten Veröffentlichungen der KMK zu den Bildungsstandards (2004, S. 7) ist niedergeschrieben, dass die Bildungsstandards „fachliche und fachübergreifende Basisqualifikationen [formulieren], die für die weitere schulische und berufliche Ausbildung von Bedeutung sind und die anschlussfähiges Lernen ermöglichen" (*claim*, siehe Abb. 2). Dieser Anspruch impliziert, dass auf Grundlage der Ergebnisse in bildungsstandardbasierten Tests zentrale Kriterien des Schulerfolgs vorhersagbar sein sollten, wie beispielsweise spätere Leistungen in Form von Testleistungen oder Schulnoten. Im Sinne von Kane wäre somit eine mögliche Schlussfolgerung: wenn Schülerinnen und Schüler bessere Leistungen im bildungsstandardbasierten Tests haben, dann sollten diese zukünftig bessere schulische Leistungen zeigen (*datum* → *claim*, siehe Abb. 2). In der vorliegenden Dissertation wurde im

Rahmen dreier Studien empirisch untersucht, inwiefern diese übergreifende Schlussfolgerung in Bezug auf die Prognosegüte bildungsstandardbasierte Tests gerechtfertigt bzw. gegeben ist.

Die Ergebnisse der Studie I stellten den Ausgangspunkt für alle weiterführenden Forschungsfragen innerhalb der nachfolgenden Studien II und III zur Prognosegüte bildungsstandardbasierter Tests auf den Schulerfolg dar. So wurde in der Studie I die noch relativ unspezifische Schlussfolgerung in Bezug auf die Prognosegüte untersucht, inwiefern es auf der Grundlage von bildungsstandardbasierte Testleistungen möglich ist, zukünftige bildungsstandardbasierte Testleistungen und Schulnoten vorherzusagen. Ausgehend von den Befunden in Studie I, wurde weiterführend in Studie II untersucht, inwiefern sich diese Schlussfolgerung in Bezug auf die Prognosegüte auf Einzelschulen generalisieren lässt und somit v. a. der Fokus auf die Untersuchung eines potenziellen *qualifier* gelegt. Im Rahmen der Studie III wurde die Prognosegüte bildungsstandardbasierter Tests in Bezug auf die Klassifikation von Personen untersucht. Auf diese Weise wurden mögliche Schlussfolgerungen auf der Grundlage bildungsstandardbasierter Testergebnisse in Hinblick auf potenzielle Handlungsentscheidungen in der Schulpraxis konkretisiert.

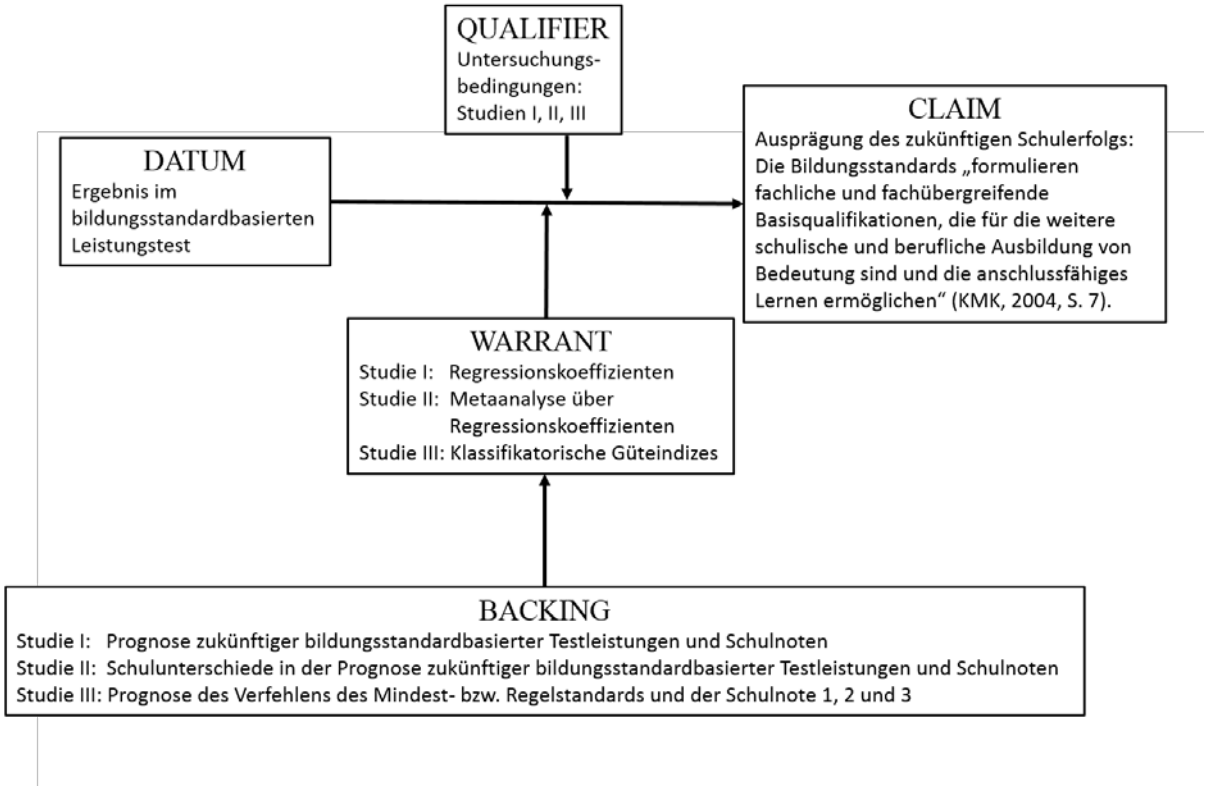


Abbildung 2. Einordnung der Studien der vorliegenden Dissertationsschrift zur Prognosegüte bildungsstandardbasierter Tests in *Toulmin's Model of Inference* (adaptierte Darstellung mit Genehmigung von Springer Nature Customer Service Centre GmbH: Springer [Standard setting: Bridging the worlds of policy making and research, Pant, Tiffin-Richards & Stanat, Abbildung 4. 2, S. 57] Copyright 2017 bei Springer auf der Grundlage der Erläuterungen von Kane (2013, 2017)

Die drei Studien zur Prognosegüte ähneln sich in vielen Aspekten ihres Forschungsdesigns, jedoch fokussieren wir auf der Grundlage unterschiedlicher statistischer Verfahren verschiedene spezifische Aspekte der Prognosegüte, um die Grenzen des diagnostischen Potenzials bildungsstandardbasierter Tests zur Prognose des Schulerfolgs untersuchen zu können.

Im Rahmen der drei Studien untersuchten wir schulische Leistungen im Fach Mathematik und Deutsch (Lesen) über verschiedene Zeitintervalle innerhalb des Schulverlaufs von der 3. bis 8. Klasse; von der Primarstufe bis in die Sekundarstufe.

Wir haben in jeder Studie stets (A) die Prognosegüte des bildungsstandardbasierten Testergebnisses als alleinige diagnostische Information zur Prognose des zukünftigen Schulerfolgs und (B) den inkrementellen prognostischen Mehrwert des bildungsstandardbasierten Testergebnisses unter Berücksichtigung zusätzlicher leistungsprädiktiver Merkmale für den zukünftigen Schulerfolg untersucht. Als weiteres leistungsprädiktives Merkmal haben wir in allen drei Studien stets die Schulnote berücksichtigt. Insbesondere die Untersuchungen zum prognostischen Mehrwert sind für uns ein wichtiger Bestandteil unserer Untersuchungen, da einerseits eine diagnostische Entscheidung stets auf einer Integration mehrerer diagnostischer Informationen beruhen sollte (Koretz, 2003) und andererseits diesbezügliche Evidenz ein wichtiges Argument darstellt, um den Aufwand für die Durchführung der bildungsstandardbasierten Tests neben den bereits bestehenden Leistungsinformationen im Schulkontext zu rechtfertigen.

Den späteren Schulerfolg haben wir in allen drei Studien auf der Grundlage späterer schulischer Leistungen im gleichen Fach in Form von späteren bildungsstandardbasierten Testergebnissen und Schulnoten definiert. In Studie I bezogen wir als weiteres Schulerfolgskriterium den Erhalt der Gymnasialempfehlung mit ein. Diesbezügliche Befunde sind ausführlich in Studie I (Anhang A) dargestellt, werden jedoch nachfolgend aufgrund der fehlenden Betrachtung in den beiden anderen Studien nicht weiterführend aufgegriffen.

3 Forschungsstand zur Prognosegüte bildungsstandardbasierter und kommerziell erhältlicher standardisierter Schulleistungstests

Bisher ist bekannt, dass inhaltlich relevante Kompetenzen bzw. das Vorwissen zu einem früheren Zeitpunkt zu den besten Prädiktoren für den zukünftigen schulischen Kompetenz- und Wissenserwerb gehören (Helmke & Weinert, 1997). Ein detaillierter Überblick zum Forschungsstand kann dem ESM 1 zur Studie I, dem ESM 1 zur Studie II und dem ESM 8 zur Studie III entnommen werden. Über den Forschungsüberblick in den einzelnen Studien hinaus gehend, möchte ich nachfolgend einen zusammenfassenden und z. T. aktualisierten Überblick über den Forschungsstand zur Prognosegüte bildungsstandardbasierter Mathematik- und Deutschtests zur Prognose von (1) zukünftigen Testleistungen desselben Faches und (2) zukünftigen Schulnoten desselben Faches, auf der Basis von Längsschnittstudien, geben. Ergänzend dazu berichte ich den aktuellen Forschungsstand, welcher für kommerziell erhältliche standardisierte Schulleistungstests zur Messung der Mathematikkompetenz und der Lesekompetenz (i. w. S.) im Fach Deutsch vorliegt. Auf diese Weise möchte ich die Forschungsdesiderate aufzeigen, die der inhaltlichen und methodischen Zielstellung unserer drei Studien zugrunde liegen.

Unseres Wissens wurde vor unseren Studien lediglich im Rahmen der Studie von Graf und Kollegen (2016) explizit die Prognosegüte bildungsstandardbasierter Tests im Rahmen von VERA untersucht und bewertet. Diese wird nachfolgend im Zusammenhang mit anderen Studien, die relevante Kennwerte liefern, erläutert und berichtet.

3.1 Prognosegüte: Korrelations- und Regressionskoeffizienten

3.1.1 Prognose zukünftiger Testleistungen

In den meisten Studien wird traditionell die Ausprägung der Prognosegüte für Testverfahren mit der Berechnung von Korrelations- und (standardisierten) Regressionskoeffizienten quantifiziert. Einen zusammenfassenden Überblick der bisherigen Befunde in Form von Korrelationskoeffizienten bietet für das Fach Mathematik Tabelle 1 und für das Fach Deutsch Tabelle 2.

Tabelle 1

Mathematik - Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Schulleistungstests

Bildungsstandardbasierte Tests										kommerziell erhältliche, standardisierte Schulleistungstests							
Quelle	Test	K	Anz	PD	PZ	PG	25% Q	75% Q	PM	Quelle	Anz	PD	PZ	PG	25 % Q	75% Q	PM
A) Kriterium Testleistung																	
Prognose ab der Primarstufe																	
Nachtigall (2018)	V3	V6	9	3	3.-6.	r_{Med} = .68	*r* = .65	*r* = .69	-	S1/S3	10	J_{Med} = 1.5	1.-4.	r_{Med} = .68	*r* = .64	*r* = .70	-
Nachtigall (2018)	V3	V8	9	5	3.-8.	r_{Med} = .64	*r* = .61	*r* = .67	-	S1/S3	1	1	1.-2.	*(r = .68)*	-	-	r_{adj} = .49[a] r_{adj} = .47[b]
Prognose ab der Sekundarstufe																	
Nachtigall (2018)	V6	V8	9	2	6.-8.	r_{Med} = .79	*r* = .78	*r* = .79	-	S2	9	M_{Med} = 1	5., 6., 8., 9., 10.	r_{Med} = .77	*r* = .70	*r* = .85	-
Nagy et al. (2017)	LV	B	1	≈ 1	9.-10.	*r* = .79 *r* = .67 (NG) *r* = .71 (GY)	-	-	-								
B) Kriterium Schulnote																	
Prognose ab der Primarstufe																	
-	-	-	-	-	-	-	-	-	-	S1	1	2	2.-4.	*r* = .64	-	-	-
										S1	1	1	2.-3.	*r* = .67	-	-	-
										S1	1	1	3.-4.	*r* = .69			r_{adj} = .46[c]
Prognose ab der Sekundarstufe																	
Graf et al. (2016)	V8	J-N	1	2	8.-10.	*r* = .44 (GY)	-	-	-		-	-	-	-	-	-	-
Graf et al. (2016)	V8	MSA	1	2	8.-10.	*r* = .56 (GY)	-	-	r_{adj} = .33[d] (GY) *sr* = .33[d] (GY)								

Anmerkungen. T = Test. K = Kriterium. Anz = Anzahl der Kennwerte. PD = Prognosedauer. PZ = Prognosezeitpunkt (Klassenstufen). PG = Prognosegüte. 25% Q = 25%-Quantil. 75% Q= 75%-Quantil. PM = Prognostischer Mehrwert. V3, V6 und V8 = VERA= Vergleichsarbeiten – für dritte, sechste und achte Klassen. r_{Med} = Median der Korrelationskoeffizienten. *r* = Korrelationskoeffizient. S1/S3 = Studie I, ESM 1 bzw. Studie III, ESM 36. S2 = Studie II, ESM 1. J_{Med} = Median der Jahre. r_{adj} = adjustierter Korrelationskoeffizient. M_{Med} = Median der Monate. ( ) = In Klammern ist die (alleinige) Prognosegüte für den jeweiligen Test angegeben, welche jedoch bereits in die Berechnung von r_{Med} einbezogen wurde. LV = Ländervergleich. B = Bildungsstandardbasierte Testaufgaben. NG = Schulen mit mehreren Bildungsgängen, Realschule, Gesamtschule. GY = Gymnasium. J-N = Jahrgangsnote. MSA = Note aus zentraler Prüfung für mittleren Schulabschluss.
sr = Semipartialkorrelationskoeffizient.
[a] Mathematiknote für die 3. Klasse kontrolliert. [b] Intelligenz für die 4. Klasse kontrolliert. [c] Mathematiknote der 3. Klasse kontrolliert. [d] Jahrgangsnote der 10. Klasse kontrolliert.

Tabelle 2

Deutsch - Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Schulleistungstests

Bildungsstandardbasierte Tests										kommerziell erhältliche, standardisierte Schulleistungstests							
Quelle	Test	K	Anz	PD	PZ	PG	25% Q	75% Q	PM	Quelle	Anz	PD	PZ	PG	25 % Q	75% Q	PM
A) Kriterium Testleistung																	
Prognose ab der Primarstufe																	
Nachtigall (2018)	V3[a]	V6[a]	9	3	3.- 6.	r_{Med} = .65	*r* = .64	*r* = .66	-	S1	5	r_{Med}=2	0.-4.	r_{Med}=.58	*r* = .46	*r* = .68	-
Nachtigall (2018)	V3[a]	V8[a]	9	5	3.- 8.	r_{Med} = .64	*r* = .63	*r* = .65	-								
Prognose ab der Sekundarstufe																	
Nachtigall (2018)	V6[a]	V8[a]	9	2	6.-8.	r_{Med} = .74	*r* = .70	*r* = .78	-								
B) Kriterium Schulnote																	
Prognose ab der Primarstufe																	
Hildebrandt &Watermann (2015)	B	Note	1	2	4.-6.	*r* = .38	-	-	r_{adj} = .18[b] (HS) r_{adj} = .19[b] (RS) r_{adj} = .17[b] (GY)	S1	1	≈ 3	0.-3.	*r* = .42[c]	-	-	-
										S1	1	≈ 2	0.-2.	*r* = .42[c]	-	-	-
										S1	1	≈ 1	0.-1.	*r* = .45[c]	-	-	-
Prognose ab der Sekundarstufe																	
Graf et al. (2016)	V8[d]	J-N	1	2	8.-10.	*r* = .23 (GY)	-	-	-								
Graf et al. (2016)		MSA	1	2	8.-10.	*r* = .35 (GY)	-	-	r_{adj} = .28[e] (GY) *sr* = .32[e] (GY)								

Anmerkungen. T = Test. K = Kriterium. Anz = Anzahl der Kennwerte. PD = Prognosedauer. PZ = Prognosezeitpunkt (Klassenstufen). PG = Prognosegüte. 25% Q = 25%-Quantil. 75% Q= 75%-Quantil. PM = Prognostischer Mehrwert. V3, V6 und V8 = VERA= Vergleichsarbeiten – für dritte, sechste und achte Klassen. r_{Med} = Median der Korrelationskoeffizienten. *r* = Korrelationskoeffizient. S1 = Studie I, ESM I. r_{adj} = adjustierter Korrelationskoeffizient. B = Bildungsstandardbasierte Testaufgaben. HS = Hauptschule. RS = Realschule. GY = Gymnasium. J-N = Jahrgangsnote. MSA = Note aus zentraler Prüfung für mittleren Schulabschluss. *sr* = Semipartialkorrelationskoeffizient.

[a] vollständiger Deutschtest (alle Leistungsdomänen) des jeweiligen VERA-Durchganges. [b] kontrolliert für Intelligenz, HISEI, Migrationshintergrund, Geschlecht, Mathematik (Test, Note), Deutschnote der 4. Klasse. [c] Prognose auf Lesenote im Fach Deutsch. [d] Leistungsdomäne Lesen im Fach Deutsch. [e] kontrolliert für die Jahrgangsnote der 10.Klasse.

Lagen mindestens fünf Korrelationswerte für den jeweiligen Forschungsstand vor, wurde für die Verteilung aller vorliegenden Korrelationskoeffizienten der Median berechnet sowie das 25 %- und das 75 %-Quantil. Zwischen diesen beiden Quantilwerten liegen somit 50 % der aus dem bisherigen Forschungsstand vorliegenden Korrelationskennwerte. Auf diese Weise erhalten wir einen empirisch fundierten Referenzbereich zur Einordnung unserer Forschungsbefunde (Bosco, Aguinis, Singh, Field & Pierce, 2015) welcher detailliert den beiden Tabellen entnommen werden kann.

Die meisten Kennwerte zur Prognosegüte bildungsstandardbasierter Tests (Tabelle 1 und 2) basieren auf den Daten der regulären Vergleichsarbeiten in Thüringen (dort bekannt als Kompetenztests) zum Zeitpunkt der 3., 6. und 8. Klassenstufe (analog zu den bundesweiten Erhebungen VERA 3, VERA 6 und VERA 8). So berichtet Nachtigall in den jährlichen Landesberichten Korrelationskennwerte zwischen früheren und späteren VERA-Testleistungen auf der Basis von Längsschnittdaten, jedoch nur z. T. mit konkreter Angabe der Kennwerte (z. B. 2014). Beginnend vom Schuljahr 2009/10 liegen sukzessive für die Schülerkohorten bis heute zum Schuljahr 2017/18 Kennwerte für insgesamt 9 Schülerkohorten vor (C. Nachtigall, persönl. Mitteilung, 22.10.2018). Auf dieser Grundlage liegen Korrelationskennwerte zur Prognosegüte von VERA-Tests auf zukünftige VERA-Tests ab der Primarstufe in der 3. auf die 6. bzw. 8. Klassenstufe der Sekundarstufe über Prognosezeiträume von 3 und 5 Jahren vor; und ab der Sekundarstufe in der 6. auf die 8. Klassenstufe über einen Prognosezeitraum von 2 Jahren. Es wurden die Leistungen der Schülerinnen und Schüler für den vollständigen Mathematik- und Deutschtest des jeweiligen Messzeitpunktes miteinander korreliert und somit Unterschiede den VERA-Tests im Testaufbau der jeweiligen VERA-Durchganges nicht berücksichtigt (bspw. Messung unterschiedlicher Leitideen in Mathematik, andere Leistungsdomänen neben Deutsch-Lesen zu den jeweiligen Testzeitpunkten; Institut zur Qualitätsentwicklung im Bildungswesen, n. d.). Des Weiteren wurden die Korrelationswerte auf der Grundlage der gesamten Längsschnittstichprobe berechnet und nicht differenziert für Schülerinnen und Schüler verschiedener Schulformen. Würde man dies tun, sollten die Korrelationskennwerte im Vergleich zu den bisher berechneten niedriger ausfallen, da die Leistungsvarianz der Schülerinnen und Schüler innerhalb der schulformspezifischen Stichprobe niedriger ist als für die Gesamtstichprobe (ohne Unterscheidung nach Schulform). Die Varianzeinschränkung hat in der Regel niedrigere Korrelationskennwerte zur Folge (Sedlmeier & Renkewitz, 2013). Auf der Grundlage der ermittelten 25 %- und 75 %-Quantile wird deutlich, dass sich die mittleren

50 % der Korrelationskoeffizienten in einen relativ engen Wertebereich verteilen – sowohl für die Mathematiktests als auch für die Deutschtests. Somit variiert die Prognosegüte der Tests zwischen den Kohorten kaum, obwohl jedes Jahr neuzusammengestellte VERA-Tests eingesetzt werden. Neben den Ergebnissen von Nachtigall gibt es lediglich einen weiteren aktuellen Studienbefund zur Prognosegüte bildungsstandardbasierter Mathematiktests auf zukünftige Mathematiktestleistungen (Nagy, Haag, Lüdtke & Köller, 2017). Hier wurden die Leistungen von Schülerinnen und Schülern auf der Basis bildungsstandardbasierter Testaufgaben (zum einen aus dem Ländervergleich) zu den beiden Messzeitpunkten der 9. und 10. Klassenstufe miteinander korreliert. Der ermittelte Wert für die Prognosegüte innerhalb der Sekundarstufe fiel für die Gesamtstichprobe mit $r = .79$ ebenso hoch aus wie der Median über die Kennwerte innerhalb der Sekundarstufe (ab der 6. Klasse) von Nachtigall (s. Tabelle 1, Kriterium Testleistung, Prognose ab Sekundarstufe; C. Nachtigall, persönl. Mitteilung, 22.10.2018). Keine der Studien hat den inkrementellen, prognostischen Mehrwert bildungsstandardbasierter Tests zur Prognose von Testleistungen untersucht.

Der vorliegende Forschungsstand zu bildungsstandardbasierten Mathematik- und Deutschtests ist umfangreicher als jener für kommerziell erhältliche, standardisierte Schulleistungstests zur Messung der mathematischen Kompetenz (rechte Seite, Tabelle 1) und jener für die Messung der Lesekompetenz (i. w. S) im Fach Deutsch (rechte Seite, Tabelle 2). Vor allem fällt auf, dass die bisherige Befundlage für standardisierte Schulleistungstests auf einer maximalen Prognosedauer von 2 Jahren basiert (meist weniger v. a. in der Sekundarstufe). Obwohl davon auszugehen ist, dass eine längeren Prognosedauer mit einer geringeren Prognosegüte einhergeht (*validity degradation*; Dahlke, Kostal, Sackett & Kuncel, 2018), ist die Prognosegüte für bildungsstandardbasierte Tests vergleichbar mit der Prognosegüte kommerziell erhältlicher, standardisierter Schulleistungstests. Obwohl aufgrund dieses Effektes zu erwarten wäre, dass diese für die bildungsstandardbasierten Tests aufgrund des längeren Prognosezeitraumes von überwiegend 2 bis 5 Jahren niedriger ausfallen. Insbesondere sei hier auf die Prognosen von Testleistungen in der Sekundarstufe für Mathematik verwiesen, deren Prognosedauer für standardisierte Schulleistungstests im Median lediglich einen Monat betrug und, bis auf eine Ausnahme, auf Retest-Reliabilitätsanalysen (identische Tests im Prädiktor und Kriterium, identische Probandinnen und Probanden) basiert: Für diese wären im Vergleich zu den Werten für bildungsstandardbasierte Tests u. a. aufgrund des Effektes der *validity degradation* deutlich höhere Werte zu erwarten. Lediglich für einen kommerziellen Mathematiktest (DEMAT 3+;

Roick, Gölitz & Hasselhorn, 2004). liegen Kennwerte zum inkrementellen prognostischen Mehrwert für die Prognose zukünftiger Testleistungen vor (bspw. neben Berücksichtigung der Mathematiknote). Für die Sekundarstufe haben wir im Rahmen unserer Studien nicht explizit nach kommerziellen Lesetests im Fach Deutsch für die Prognose von Testleistungen recherchiert, da der Fokus im Rahmen der Studien im Fach Deutsch nicht explizit die Sekundarstufe umfasste und haben in Studie II auf den Forschungstand von Graf (2016) zurückgegriffen, welcher für kommerzielle Tests ausnahmslos auf Querschnittsstudien basiert und somit in der Übersicht nicht berücksichtigt wurde.

3.1.2 Prognose zukünftiger Schulnoten

Der Forschungsstand bildungsstandardbasierter Tests zur Prognose von zukünftigen Schulnoten basiert nur auf wenigen Studien. Für Mathematiktests in der Primarstufe sind uns keine Studien bekannt. Lediglich mit der Studie von Graf und Kollegen (2016) lag eine erste Studie im Verlauf der Sekundarstufe vor, die explizit die Güte von VERA-Tests für die Prognose zukünftiger Schulnoten untersuchte. Die Ergebnisse innerhalb der Sekundarstufe für Gymnasiasten zeigten, dass u. a. VERA-8-Tests in Mathematik und Deutsch (Lesen) in der 8. Klasse die 2 Jahre spätere Prüfungsnote für den mittleren Schulabschluss und die Mathematikjahrgangsnote für dasselbe Fach in der 10. Klasse substanziell vorhersagen können. Darüber hinaus konnte in dieser Studie ein inkrementeller, prognostischer Mehrwert der VERA-Tests auf die späteren Prüfungsnoten in der 10. Klasse aufgezeigt werden, wenn für die Jahrgangsnote der 10. Klasse kontrolliert wurde. Korrelationskennwerte aus der Studie von Hildebrandt und Watermann (2015) zur Prognosegüte bildungsstandardbasierter Deutschaufgaben auf die Endjahresnote im Fach Deutsch fallen ähnlich wie die Befunde von Graf (2016) aus. Für den prognostischen Mehrwert sind die adjustierten Korrelationskoeffizienten für Gymnasiasten mit $r = .17$ jedoch niedriger als bei Graf (2016) mit $r = .28$. In der Studie von Hildebrandt und Watermann (2015) wurde jedoch für zahlreiche leistungsrelevante Merkmale (s. Tabellenanmerkung) kontrolliert, sodass dies systematisch den prognostischen Mehrwert reduziert.

Auch die Befunde zur Prognosegüte kommerziell erhältlicher, standardisierter Schulleistungstests in Mathematik und Deutsch (Lesen) auf zukünftige Schulnoten ist vergleichsweise gering. In Mathematik liegen lediglich längsschnittliche Ergebnisse für die Primarstufe vor. Erneut liegen für den DEMAT 3+ (Roick et al., 2004), als einzigen kommerziell erhältlichen standardisierten Schulleistungstest, Kennwerte zum inkrementellen

prognostischen Mehrwert vor. Hier wurde als zusätzlicher Prädiktor die Mathematiknote einbezogen. Erneut berichten wir für die Sekundarstufe keine Werte für kommerziell erhältliche Tests im Lesen, da der Fokus im Rahmen der Studien im Fach Deutsch nicht explizit die Sekundarstufe umfasste (s. ebenso im Abschnitt zur „Prognosegüte zukünftiger Testleistungen“).

3.2 Prognosegüte: Klassifikatorische Güteindizes

Vor unseren Studien zur klassifikatorischen Prognosegüte bildungsstandardbasierter Tests gab es unseres Wissens keine Studien, die diese explizit für das Verfehlen bestimmter Bildungsergebnisse auf der Grundlage späterer Testleistungen und Schulnoten untersuchten. Jedoch konnte auf der Grundlage der berichteten Ergebnisse von Graf und Kollegen (s. ESM 5, 2016), die Klassifikationsgüte von VERA-8-Tests hinsichtlich der 2 Jahre später erreichten Noten in den zentralen Prüfungen zum Mittleren Schulabschluss geschätzt werden, allerdings sind die gewählten Schwellenwerte abweichend zu denen in unserer Studie und somit die Befunde nicht adäquat vergleichbar (Details z. Forschungsstand s. ESM 1der Studie III).

Für kommerziell erhältliche Tests liegen hingegen z. T. deutlich mehr Befunde zur klassifikatorischen Prognosegüte von Mathematik- und Deutschtests in Bezug auf spätere Testleitungen oder Schulnoten vor (s. Forschungsüberblick in ESM 1 der Studie III für eine separate Darstellung der klassifikatorischen Güte für kommerzielle und nicht kommerzielle Tests; zusammenfassende Darstellung in ESM 36-1 der Studie III). Dieser ist somit deutlich umfangreicher als für den zuvor berichteten Forschungsstand auf der Grundlage von Korrelations- und Regressionskoeffizienten. Dies ist plausibel, wenn man das primäre diagnostische Ziel des Einsatzes kommerziell erhältlicher Leistungstests berücksichtigt, welches vor allem darauf ausgerichtet ist, Informationen zur Unterstützung individualdiagnostischer Entscheidungen über den Förderbedarf von Schülerinnen und Schülern zu sammeln (Köller & Reiss, 2013). Zur Untersuchung dieser Zielstellung sind Korrelations- und Regressionsanalysen – die Zusammenhänge zwischen Prädiktoren und Kriterien auf der Grundlage des gesamten kontinuierlichen Merkmalsspektrum quantifizieren – nicht hinreichend (Gersten et al., 2012; Marx, 1992). Im Rahmen der Individualdiagnostik (z. B. zur Identifikation von Förderbedarf) ist es jedoch erforderlich zu entscheiden, ob eine Person einen bestimmten Schwellenwert überschritten hat oder nicht, bzw. ein Kriterium erreicht hat oder nicht. Die Individualdiagnostik ist damit eng an die Klassifikation von Personen gebunden. Die Klassifikationsgüte von Tests kann hierzu mit zahlreichen

Kennwerten, wie zum Beispiel der Sensitivität (s. a. den Abschnitt „Statistische Analysemodelle“ der Studie III), bewertet werden (u. a. Marx, 1992; Gersten et al., 2012). Überwiegend konzentrierten sich bisherige Studien auf die Berechnung der Sensitivität (s. Darstellung in ESM 36-1 der Studie III) und vernachlässigten insbesondere einen weiteren wichtigen Kennwert zur Bewertung des Potenzials von Tests zur Identifizierung von Schülerinnen und Schülern mit Förderbedarf, wie den positiv prädiktiven Wert. Gleichzeitig weist der bisherige Forschungsstand zu kommerziellen Tests gewisse Einschränkungen auf, die den Vergleich mit unseren Befunden erschweren. So fokussieren sich bisher die meisten Studien zu kommerziellen Tests auf Prognosen innerhalb der Primarstufe. Des Weiteren basieren die Schwellenwerte zur Klassifikation der Schülerinnen und Schüler bisher ausnahmslos auf sozialen Bezugsnormen. Hinzu kommt, dass lediglich für 2 von 19 standardisierten Leistungstests überprüft wurde, inwiefern sich die Klassifikationsgüte durch Kombination mehrerer Leistungsprädiktoren verbessert. Obwohl im Rahmen einer validen Individualdiagnostik mehrere Informationen einbezogen werden sollten für eine diagnostische Entscheidung (Koretz, 2003), liegt hier somit ein deutliches Forschungsdesiderat vor.

3.3 Forschungsdesiderate

Die Studie von Graf und Kollegen (2016) war bis zum Beginn des vorliegenden Dissertationsprojektes, die einzige, in der explizit die Prognosegüte bildungsstandardbasierter Tests untersucht wurde. Jene Ergebnisse beschränken sich auf die Güte in Bezug auf spätere Prüfungs- und Jahrgangsnoten. Als Besonderheit im Vergleich zu den damaligen Vorgängerstudien zur Prognosegüte für kommerzielle, standardisierte Schulleistungstests im deutschsprachigen Raum wurde zudem der inkrementelle, prognostische Mehrwert für bildungsstandardbasierte Tests unter Berücksichtigung der Prüfungsnote berichtet. Zusammenfassend konnte u. a. im Rahmen der Studie für VERA-8-Tests zur Messung der mathematischen Kompetenz und für Tests zur Messung der Lesekompetenz im Fach Deutsch festgestellt werden, dass diese zur Prognose 2 Jahre späterer Schulnoten eine vergleichbare Prognosegüte zu bisher etablierten standardisierten Schulleistungstests aufweisen. Dieses Ergebnis ist besonders interessant, da der bisherige Forschungsstand in der Sekundarstufe für standardisierte Schulleistungstests hauptsächlich auf Querschnittsdaten basiert und hier aufgrund der *validity degredation* systematisch höhere Korrelations- und Regressionswerte als in Längsschnittstudien zu erwarten sind (Dahlke et al., 2018; s. ESM 1 unter Berücksichtigung der Fußnoten in Graf et al., 2016). Die genannten Studienbefunde

beschränken sich auf Schülerinnen und Schüler an Gymnasien. Weitere berichtete Korrelationswerte aus anderen Studien (bspw. von C. Nachtigall, persönl. Mitteilung, 22.10.2018), auf der Grundlage von bildungsstandardbasierten Tests die ebenso als Maße der Prognosegüte interpretiert werden können, untermauern die Befunde von Graf (Graf et al., 2016). Diese zeigen insbesondere, dass die Prognosegüte bildungsstandardbasierter Tests für spätere Testleistungen ebenso vergleichbar ist mit standardisierten Schulleistungstests – selbst für Prognosezeiträume zwischen 2 bis 5 Jahren. Diese variieren zudem kaum zwischen den 9 Schülerkohorten, in denen verschiedene VERA-Testheftzusammenstellungen über die verschiedenen Schuljahre realisiert wurden.

Diese Befundlage zur Prognosegüte bildungsstandardbasierter Tests ist als unzureichend zu bewerten, wenn man bedenkt, dass jährlich im Rahmen von VERA ca. 1.4 Millionen Schülerinnen und Schüler in Deutschland solche Tests bearbeiten (KMK, 2012; Statistisches Bundesamt, 2017). So sollte ausreichend sichergestellt werden, inwiefern Schlussfolgerungen auf der Grundlage von bildungsstandardbasierten Tests valide pädagogisch relevante Interpretationen, Schlussfolgerungen und Entscheidungen zulassen (Kane, 2009, 1992; Messick, 1995). Dieser Anspruch sollte nicht nur an sogenannte *high-stakes*-Tests gestellt werden (Decker & Bolt, 2008), sondern sollte auch für VERA-Tests gelten, die von Wissenschaftlerinnen und Wissenschaftlern auf der individuellen Ebene für Schülerinnen und Schüler als *no-stakes* und auf der Ebene der Lehrkräfte und Schulleitungen als *low-stakes* eingeschätzt werden (u.a. Pant et al., 2017).

Unabhängig von der dürftigen Befundlage zur Prognosegüte bildungsstandardbasierter Tests berichten etwa die Hälfte aller Lehrkräfte in Befragungen, dass ihnen die VERA-Tests Informationen für die Planung individueller Fördermaßnahmen liefern (Richter & Böhme, 2014; Wurster et al., 2017). Auf wissenschaftlicher Ebene wird dieser Aspekt kritisch betrachtet und die Güte von VERA zur Individualdiagnostik allenfalls im Sinne eines Screenings eingeräumt – jedoch wurde die Güte der Tests in Hinblick auf diese Funktion bisher nicht empirisch untersucht (Köller & Reiss, 2013; Leutner et al., 2008). Dies ist auf der Grundlage klassifikatorischer Analysen möglich. Für der Sichtung des diesbezüglichen Forschungsstandes kommerziell erhältlicher Schulleistungstests zeigte sich, dass hier bisher wichtige Güteindizes, wie der positiv prädiktive Wert, unberücksichtigt blieben. Diese Lücke haben wir geschlossen (nicht nur in Hinblick auf den positiv prädiktiven Wert) indem wir für die bis dato vorliegenden Studien zahlreiche klassifikatorische Güteindizes auf der Grundlage

der berichteten Angaben innerhalb der Studien berechnet haben (s. ESM 1 der Studie III). Ebenso wurde die Klassifikationsgüte selten im Zusammenhang mit weiteren Leistungsinformationen bestimmt.

Nicht nur der Forschungsstand zu bildungsstandardbasierten, sondern auch jener für kommerzielle standardisierte Schulleistungstests ist vor allem in Bezug auf den inkrementellen prognostischen Mehrwert als unzureichend zu bewerten. Jedoch könnte die diesbezügliche Evidenz sicherlich ein wichtiges praxisrelevantes Argument für Lehrkräfte sein, um den Aufwand für kommerzielle Tests – sowie den Aufwand für bildungsstandardbasierte Tests – zu rechtfertigen. So berichten Lehrkräfte in vielen Studien im Zusammenhang mit VERA von einem hohen Aufwand und schätzen den Nutzen als gering ein (Kuper, Maier, Graf, Muslic & Ramsteck, 2017; Wurster et al., 2017).

4 Forschungsbeitrag zur Prognosegüte bildungsstandardbasierter Tests

4.1 Inhaltliche und methodische Zielsetzung der drei Studien

Nachfolgend stelle ich die drei Studien vor, die im Rahmen der vorliegenden Dissertationsschrift entstanden sind. In allen drei Längsschnittstudien wurde die übergreifende Forschungsfrage untersucht, inwiefern bildungsstandardbasierte Tests geeignet sind, den zukünftigen schulischen Erfolg von Schülerinnen und Schülern – fokussiert auf spätere Schulnoten und bildungsstandardbasierte Testleistungen im selben Fach – vorherzusagen. Dass dieser Anspruch an die Tests gestellt wird, habe ich im Rahmen eines argumentbasierten Ansatzes (2013, 2017) erläutert. Dieser Anspruch an die Prognosegüte bildungsstandardbasierter Tests lässt sich zum einen auf der Grundlage der Veröffentlichungen der KMK zu den Bildungsstandards und zum anderen durch die festgelegten Testzeitpunkte im Rahmen der VERA-Testungen begründen (s. Abschnitt 2.2). In allen drei Studien untersuchten wir neben der alleinigen Güte der Tests zur Prognose zukünftiger Testleistungen und Schulnoten den inkrementellen prognostischen Mehrwert der Tests unter der Berücksichtigung weiterer leistungsprädiktiver Merkmale, wie beispielsweise den Schulnoten.

Die Befunde der Studie I zur Prognosegüte bildungsstandardbasierter Tests stellten den Ausgangspunkt für die jeweiligen weiterführenden Forschungsfragen zur Prognosegüte im Rahmen der Studie II und III dar. So wurde weiterführend in Studie II untersucht, inwiefern sich die zuvor ermittelte Prognosegüte auf der Basis der Gesamtstichprobe der Schülerinnen und Schüler auf die Schülerschaft an Einzelschulen generalisieren lässt. Im Rahmen der Studie III wurde die Prognosegüte bildungsstandardbasierter Tests in Bezug auf die Klassifikation von Personen untersucht (ob diese einen bestimmten Schwellenwert überschreiten oder nicht) und nicht in Bezug auf das gesamte kontinuierliche Merkmalsspektrums wie in den Studien I und II. Auf diese Weise wurde die Prognosegüte in Bezug auf die Ableitung direkter Handlungsentscheidungen in der Schulpraxis konkretisiert.

Zusammenfassend haben wir folglich untersucht (1) wie gut die (alleinige) Prognosegüte bildungsstandbasierter Tests für die Prognose zukünftiger bildungsstandardbasierter Testleistungen und Schulnoten ausgeprägt ist, (2) welchen inkrementellen prognostischen Mehrwert die bildungsstandardbasierten Tests neben der Berücksichtigung weiterer

leistungsprädiktiver Merkmale für die Prognose zukünftiger Testleistungen und Schulnoten besitzen, (3) inwiefern sich zwischen Schulen die Prognosegüte und der inkrementelle Mehrwert bildungsstandardbasierter Tests zur Prognose von zukünftigen Testleistungen und Schulnoten unterscheidet und (4) inwiefern bildungsstandardbasierte Tests im Sinne eines Screenings zur Identifizierung von Schülerinnen und Schülern geeignet sind, die gefährdet sind, wichtige Bildungsergebnisse in Bezug auf zukünftige Testleistungen und Schulnoten zu verfehlen.

Die Prognosegüte wollten wir im Rahmen der drei Studien auf einer soliden Datenbasis an Längsschnittstudien untersuchen. Dabei strebten wir an, die Prognosegüte über einen möglichst langen Prognosezeitraum zu bestimmen. Im Sinne einer integrativen und kumulativen Forschungspraxis (Cumming & Calin-Jageman, 2017), haben wir stets Kennwerte zur Bewertung der Prognosegüte berechnet, die zum einen eine Einordnung der Ausprägung der Prognosegüte in den bisherigen und zukünftigen Forschungsstand unterstützen (bspw. standardisierte Kennwerte, Konfidenzintervalle) und zum anderen eine Bewertung in Bezug auf die praktische Relevanz der Prognosegüte fördern (Berücksichtigung weiterer leistungsprädiktiver Merkmale, klassifikatorische Analysen). Diese Bestrebungen stehen ebenso in Einklang mit einigen Ausführungen von Kane (2013), welcher für eine adäquate Validierung der Testwertinterpretation bzw. -nutzung von Leistungstests nicht nur eine Bewertung der Ausprägung der ermittelten Kennwerte fordert, sondern auch die Berücksichtigung möglicher Konsequenzen in der Praxis betont.

4.2 Zusammenfassung der drei Studien

Zur Untersuchung unserer übergeordneten Forschungsfrage zur Prognosegüte bildungsstandardbasierter Test für zukünftige Testleistungen und Schulnoten desselben Faches haben wir stets Längsschnittdaten genutzt. Auf deren Basis konnten wir die Prognosegüte bildungsstandardbasierter Tests über einen Zeitraum von mindestens 1 Jahr bis hin zu 5 Jahren innerhalb der Schulzeit untersuchen. Die Prognose erfolgte z. T. ab dem Zeitpunkt der 3. oder 4. Klassenstufe der Primarstufe, sowie der 6. Klassenstufe der Sekundarstufe. Auf Basis dieser Werte wurde der Schulerfolg zum Zeitpunkt der 4., 5. und 6. Klassenstufe der Primarstufe, sowie der 8. Klassenstufe der Sekundarstufe prognostiziert. In allen drei Studien wurde die Prognosegüte von bildungsstandardbasierten Tests im Fach Mathematik untersucht; in Studie I und Studie III zusätzlich für bildungsstandardbasierte Lesetests im Fach Deutsch.

Im Nachfolgenden werden die drei Studien zusammenfassend in Bezug auf die Forschungsfragen, die angewandten Methoden und die Ergebnisse berichtet. Zudem wird erläutert, wie sich die Forschungsfragen und Befunde der jeweiligen Studien ergänzen. Die ausführlichen Manuskripte der jeweiligen Studien (inklusive der elektronischen Supplemente, [ESM]) können dem Anhang entnommen werden.

4.2.1 Studie I

Im Rahmen von Studie I haben wir auf der Basis von Regressionsanalysen konkret untersucht, wie gut bildungsstandardbasierte Tests im Fach Mathematik und im Fach Deutsch (Lesen) spätere bildungsstandardbasierte Testleistungen und Schulnoten desselben Faches vorhersagen. Darüber hinaus analysierten wir alleinig im Rahmen der Studie I wie hoch die gemeinsame Güte von bildungsstandardbasierten Testleistungen in Mathematik und Deutsch für die Prognose der Gymnasialempfehlung ist. (Die Ergebnisse zur Prognose der Gymnasialempfehlung können dem Manuskript entnommen werden, werden jedoch hier nicht weiter erläutert, da dieses Kriterium nur speziell innerhalb von Studie I untersucht wurde.).

Die Analysen basieren auf einem Längsschnittdatensatz, der sich über die 3. bis 6. Klassenstufe innerhalb der Primarstufe erstreckt. Wir berechneten für das Fach Mathematik Prognosen von der 3. Klassenstufe auf die jeweils 4., 5. und 6. Klassenstufe sowie von der 4. Klassenstufe auf die 5. und 6. Klassenstufe. Im Fach Deutsch berechneten wir ebenso Prognosen von der 4. Klassenstufe auf die jeweils 5. und 6. Klassenstufe. Die Prognosen ab der 3. bzw. 4. Klassenstufe basierten in Mathematik auf einem kommerziell erhältlichen, bildungsstandardbasierten Mathematiktest und in Deutsch auf einer Aufgabenzusammenstellung aus einem Aufgabenpool, der vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) zur Überprüfung der Bildungsstandards in der Primarstufe entwickelt wurde.

Es wurde zum einen die (alleinige) Prognosegüte der Tests (β) und zum anderen der inkrementelle prognostische Mehrwert der bildungsstandardbasierten Tests (β_{adj}, *sr*) unter Berücksichtigung weiterer leistungsprädiktiver Merkmale, wie der Schulnote desselben Faches, des sozioökonomischen Hintergrundes und der Intelligenz bestimmt.

Die Prognosegüte der Mathematiktests in der 3. und 4. Klassenstufe auf bis zu 3 Jahre spätere Mathematiktests lag für die standardisierten Regressionskoeffizienten zwischen $.56 \leq \beta \leq .66$,

für bis zu 3 Jahre spätere Mathematiknoten zwischen $.48 \leq \beta \leq .57$[1]. Positive Koeffizienten bedeuten, dass bessere Testleistungen in der 3. bzw. 4. Klassenstufe mit besseren späteren Testleistungen und Schulnoten einhergehen. Der inkrementelle prognostische Mehrwert der Mathematiktests auf spätere Mathematiktests lag für den standardisierten Regressionskoeffizienten zwischen $.30 \leq \beta_{adj} \leq .45$ bzw. für den Semipartialkorrelationskoeffizienten zwischen $.24 \leq sr \leq .33$, auf spätere Mathematiknoten zwischen $.12 \leq \beta_{adj} \leq .41$ bzw. $.10 \leq sr \leq .37$.

Für die Deutschtests (Lesen) konnte eine Prognosegüte von der 4. Klassenstufe auf die 1 Jahr bzw. 2 Jahre späteren Lesetests im Fach Deutsch von $\beta \leq .60$ bzw. $\beta \leq .61$ und auf Deutschnoten von $\beta \leq .47$ bzw. $\beta \leq .50$ aufgezeigt werden. Der inkrementelle prognostische Mehrwert der Lesetests auf die späteren Testleistungen lag bei $\beta_{adj} \leq .44$ bzw. $\beta_{adj} \leq .43$ und für beide Prognosezeiträume bei $sr \leq .36$; auf die späteren Schulnoten für beide Prognosezeiträume bei $\beta \leq .19$, sowie bei $sr \leq .10$ bzw. $sr \leq .14$.

Zusammenfassend war die Prognosegüte der bildungsstandardbasierten Tests vergleichbar mit dem bisherigen Forschungsstand zu kommerziell erhältlichen, standardisierten Schulleistungstests. Zudem entkräften die Ergebnisse zum prognostischen Mehrwert die Kritik an schulischen Leistungstests, wonach diese lediglich den sozioökonomischen Status (Sackett, Kuncel, Arneson, Cooper & Waters, 2009) oder Intelligenz widerspiegeln würden (Baumert, Lüdtke, Trautwein & Brunner, 2009). Des Weiteren zeigte sich, dass bildungsstandardbasierte Tests ebenso über Schulnoten hinaus, diagnostisches Potenzial besitzen.

Die Ergebnisse der Studie I stellten den Ausgangspunkt für alle weiterführenden Forschungsfragen zur Prognosegüte bildungsstandardbasierter Tests auf den Schulerfolg innerhalb der nachfolgenden Studien II und III dar.

4.2.2 Studie II

Im Rahmen der Studie II untersuchten wir die Prognosegüte für VERA-Mathematiktests auf der Grundlage eines Längsschnittdatensatzes mit Erhebungszeitpunkten von der 6. bis zur 8.

[1] Im Rahmen dieser Dissertationsschrift werden zur besseren Vergleichbarkeit und Integration der berechneten Regressions- bzw. Korrelationswerte über alle drei Studien hinweg hier - abweichend zur Darstellung im Manuskriptanhang der Studie I - die Regressions- bzw. Korrelationskoeffizienten auf der Grundlage rekodierter Schulnoten in 1 für *ungenügend* bis 6 für *sehr gut* angegeben. D. h. Höhere Testleistungen gehen mit besseren späteren Schulnoten einher.

Klassenstufe innerhalb der Sekundarstufe. Wir haben die Generalisierbarkeit der Prognosegüte zwischen Einzelschulen separat für Schulen mit mehreren Bildungsgängen (= SMBG) und für Gymnasien (= GY) untersucht.

In Studie I ermittelten wir die (alleinige) Prognosegüte und den inkrementellen prognostischen Mehrwert bildungsstandardbasierter Tests stets auf der Grundlage der Leistungsergebnisse aller Schülerinnen und Schüler der Gesamtstichprobe einer jeweiligen Schulform. Dies ist ein typisches Vorgehen in Studien zur Untersuchung der psychometrischen Güte von Testverfahren. Dies gilt z. B. ebenso für alle Kennwerte, die wir im Rahmen des bisherigen Forschungsstandes zur Prognosegüte für standardisierte und konkret bildungsstandardbasierte Leistungstest zusammengetragen haben (s. Abschnitt 3 „Forschungsstand zur Prognosegüte bildungsstandardbasierter und kommerziell erhältlicher standardisierte Schulleistungstests“) . Es wird jedoch nicht hinterfragt, inwiefern sich die Prognosegüte, die für die Gesamtstichprobe ermittelt wurde, auf die Schülerschaft an den jeweiligen Einzelschulen generalisieren lässt. Oder anders formuliert: Inwiefern es Unterschiede zwischen Schulen in der Prognosegüte bildungsstandardbasierter Tests gibt. Auch wenn die *Standards for Educational Testing* (AERA et al., 2014) diesbezügliche Evidenz zur Validitätsgeneralisierung explizit einfordern, wurden in der empirischen Bildungsforschung bisher Fragen zur Validitätsgeneralisierung von Leistungstests in Bezug auf Einzelschulen noch nie gestellt. Im Rahmen der Studie II untersuchten wir daher, inwiefern sich (1) die (alleinige) Prognosegüte und (2) der inkrementelle prognostische Mehrwert bildungsstandardbasierter Mathematiktests gegenüber Schulnoten über einen Prognosezeitraum von 2 Jahren über Einzelschulen generalisieren lässt bzw. variiert. Darüber hinaus analysierten wir, inwiefern sich Unterschiede zwischen Schulen in der Prognosegüte durch schulspezifische Merkmale, die auf der Grundlage der Testergebnisse gebildet wurden (Reliabilität der Tests, Leistungsheterogenität im Test, Leistungsniveau im Test) erklären lassen. Zur Untersuchung nutzen wir ein meta-analytisches Vorgehen nach Cheung und Jak (2016).

Die Befunde zeigten, dass sich (1) die alleinige Prognosegüte und (2) der inkrementelle prognostische Mehrwert gegenüber Schulnoten z. T. nur eingeschränkt zwischen Einzelschulen von SMBG und Einzelschulen von GY generalisieren lassen. Es zeigen sich differentielle Befunde in der Ausprägung der Generalisierbarkeit der Prognosegüte in Abhängigkeit des berechneten statistischen Kennwerts zur Quantifizierung der Prognosegüte

(β) bzw. des inkrementellen Mehrwerts (β_{adj}, *sr*), sowie der Schulform (SMBG; Gymnasien) und des prognostizierten Kriteriums (Testleistung, Schulnote). Trotz der z. T. vorliegenden Heterogenität der Prognosegüte zwischen den Einzelschulen liegt (1) die eine substantielle Prognosegüte ($\beta = 0$) auf 2 Jahre spätere Noten und Testleistungen für mindestens 93 % der SMBG und 100 % der GY vor (s. Studie II, ESM 3). Zudem (2) lag ein substantieller inkrementeller prognostischer Mehrwert ($sr/\beta_{adj} = 0$) gegenüber Schulnoten auf spätere Noten für mindestens 81 % der SMBG und 100 % der GY und für die Prognose von Testleistungen für mindestens 60 % der SMBG und 100 % der GY vor (s. Studie II, ESM 3). Die vorliegende Heterogenität zwischen den Schulen in der Prognosegüte und dem inkrementellen Mehrwert konnte z. T. durch schulspezifische Merkmale erklärt werden. So konnte unter Berücksichtigung der schulspezifischen Reliabilität und der Leistungsheterogenität in der 8. Klassenstufe die Heterogenität der Prognosegüte auf spätere Testleistungen zwischen SMBG auf ein nicht statistisch signifikantes Niveau reduziert werden. Schulunterschiede zwischen SMBG in der Prognosegüte auf spätere Schulnoten und die meisten Kennwerte des inkrementellen Mehrwertes für die Prognose späterer Schulnoten und Testleistungen blieben jedoch bestehen. Zwischen GY zeigte sich bereits ohne Berücksichtigung schulspezifischer Merkmale für die Prognosegüte auf Schulnoten und den inkrementellen Mehrwert auf Schulnoten und Testleistungen keine statistisch signifikante Heterogenität. Die bestehende Heterogenität zwischen GY für die Prognosegüte auf spätere Testleistungen und den inkrementellen Mehrwert in Bezug auf spätere Testleitungen ließ sich jeweils durch das schulspezifische Leistungsniveau in der 6. Klassenstufe, die Leistungsheterogenität in der 8. Klassenstufe und die Reliabilität in der 8. Klassenstufe in Bezug auf ein statistisch nicht signifikantes Niveau reduzieren. Zusammenfassend zeigten sich erwartungskonforme Effekte der schulspezifischen Reliabilität und der Leistungsheterogenität auf die Prognosegüte: So ließ sich für SMBG und GY mit höherer Reliabilität und Leistungsheterogenität in der 8. Klassenstufe eine höhere Prognosegüte und ein höherer inkrementeller Mehrwert zur Prognose der späteren Testleistung und Schulnoten aufzeigen.

Insgesamt liefert unsere Studie erstmalige empirische Evidenz dafür, dass VERA-Mathematiktests spätere Testleistungen und Schulnoten an der Mehrzahl der Schulenfür SMBG und GY prognostizieren können. Zudem weisen die Befunde darauf hin, dass die Tests an den meisten Schulen einen (z. T. deutlichen) prognostischen Mehrwert für die Leistungsdiagnostik neben Zeugnisnoten haben können. Zudem weisen die Befunde zum schulspezifische Leistungsniveau darauf hin, dass Einschränkungen in der

Validitätsgeneralisierung zumindest teilweise kompensiert werden können, wenn die Reliabilität der Test erhöht wird, indem das Schwierigkeitsniveau der Testitems besser an das Leistungsniveau der Schülerinnen und Schülerangepasst wird. Diese Befunde unterfüttert somit empirisch aktuelle Bestrebungen der KMK in Bezug auf die Weiterentwicklung von VERA (KMK, 2012).

Ansonsten fielen die mittleren (aggregierten) Kennwerte der Prognosegüte über die verschiedenen Einzelschulen von SMBG bzw. GY in ihrer Ausprägung vergleichbar mit dem bisherigen Forschungsstand (sofern vorliegend) zu standardisierten Schulleistungstests der Primar- und Sekundarstufe aus, welche auf der Grundlage der Gesamtstichproben (ohne Berücksichtigung der Einzelschulen) berechnet wurden. Folglich bestätigen die Ergebnisse der Studie II die Schlussfolgerungen der Studie I und ergeben somit ein kohärentes Bild zur Prognosegüte bildungsstandardbasierter Tests.

4.2.3 Studie III

Konkret untersuchten wir im Rahmen der Studie III, inwiefern bildungsstandardbasierte Tests – im Sinne eines Screenings – zur Identifizierung von Schülerinnen und Schülern geeignet sind, die gefährdet sind, wichtige Bildungsergebnisse in Bezug auf zukünftige Testleistungen und Schulnoten zu verfehlen. Mit dieser Frage nach der klassifikatorischen Prognosegüte bildungsstandardbasierter Tests nähern wir uns stärker diagnostischen Fragestellungen im Rahmen der Individualdiagnostik an. Auf dieser Grundlage kann abgeschätzt werden, inwiefern die diagnostischen Informationen aus bildungsstandardbasierten Tests mögliche praxisrelevante Entscheidungen beeinflussen könnten. Auf diese Weise können wir im Sinne von Kane – im Vergleich zur Studie I und Studie II – stärker mögliche Konsequenzen in die Untersuchung der Prognosegüte bildungsstandardbasierter Tests einfließen lassen, die auf der Grundlage von Testwertinterpretation getroffen werden könnten und somit potenzielle datenbasierte Entscheidungen evaluieren, die auch auf der Grundlage bildungsstandardbasierter Tests getroffen werden könnten. Bisher wurde lediglich vereinzelt aus wissenschaftlicher Sicht das Potenzial bildungsstandardbasierter Tests für diese individualdiagnostische Zielstellung im Sinne eines Screenings für die weiterführende Diagnostik angedeutet (Köller & Reiss, 2013; Leutner et al., 2008), jedoch nicht für bildungsstandardbasierte Tests untersucht. Unsere bisherigen Befunde auf der Basis von Regressions- und Korrelationsanalysen der Studie I und Studie II sind nicht hinreichend, um

die Prognosegüte in Hinblick auf diese Screeningfunktion zu beurteilen (s. Abschnitt 3.2 Prognosegüte: Klassifikatorische Güteindizes oder Studie III).

Im Rahmen der Individualdiagnostik (bspw. zur Identifikation von Förderbedarf) ist es grundsätzlich erforderlich zu entscheiden, ob eine Schülerin oder ein Schüler einen bestimmten Wert (= Schwellenwert) überschritten hat oder nicht, bzw. ein Kriterium erreicht hat oder nicht. Folglich ist die Individualdiagnostik eng an die Klassifikation von Personen gebunden. Die Güte der Klassifikation von Tests kann hierzu mit zahlreichen Kennwerten, wie bspw. der Sensitivität, bewertet werden. Die Rationale für die Festlegung der notwendigen Schwellenwerte zur Identifikation bzw. Klassifikation von Schülerinnen und Schüler in Bezug auf bildungsstandardbasierte Tests können die Beschreibungen der KMK im Rahmen der veröffentlichten Kompetenzstufenmodelle z. B. für Mathematik und Deutsch (KMK, 2013a, 2013b, 2013c, 2014) liefern.

Die Analysen in Studie III wurden für zwei Längsschnittdatensätze durchgeführt. In der fokussierten Hauptstudie wurde die Klassifikationsgüte von VERA-Mathematiktests und von VERA-Lesetests im Fach Deutsch für einen Prognosezeitraum ab der 3. Klassenstufe in der Primarstufe auf die 8. Klassenstufe in der Sekundarstufe betrachtet. Die Klassifikationsgüte wurde separat für Schülerinnen und Schüler bestimmt, die in der Sekundarstufe Schulen mit mehreren Bildungsgängen (= SMBG) und Gymnasien (= GY) besuchten. Im Rahmen der Ergänzungsstudie wurden die Analysen zur Klassifikationsgüte für einen kommerziell erhältlichen bildungsstandardbasierten Mathematiktest für Prognosen innerhalb der Primarstufe von der 4. auf die 5. Klassenstufe und der 4. auf die 6. Klassenstufe repliziert.

Zur Bewertung der Klassifikationsgüte zur Identifikation von Kindern, die „gefährdet“ sind, zukünftige Bildungsergebnisse zu verfehlen, fokussierten wir zwei einschlägige Indizes: Sensitivität und positiv prädiktiver Wert (Bossuyt et al., 2015; Petscher, Kim & Foorman, 2011). Bemessen an der Gesamtanzahl aller Schülerinnen und Schüler, die spätere Bildungsergebnisse tatsächlich verfehlten, gibt die Sensitivität den prozentualen Anteil von Schülerinnen und Schülern an, die richtig als „gefährdet“ identifiziert wurden: *Sen* = RP / (RP + FN) mit RP als der Anzahl Richtig Positiver und FN als der Anzahl Falsch Negativer. Bemessen an der Gesamtanzahl aller Schülerinnen und Schülern, die in der Grundschule als „gefährdet“ klassifiziert wurden, gibt der positiv prädiktive Wert (*ppW*) den prozentualen Anteil von Schülerinnen und Schülern an, die die späteren Bildungsergebnisse tatsächlich verfehlten: *ppW* = RP / (RP + FP) mit FP als der Anzahl Falsch Positiver.

Die Klassifikationsgüte für die bildungsstandardbasierten Tests ermittelten wir in Bezug auf zwei Schwellenwerte zur Identifikation „gefährdeter“ Kinder zum ersten Messzeitpunkt: Zum einen klassifizierten wir alle Kinder als „gefährdet“ spätere Bildungsergebnisse zu verfehlen, wenn diese (1) den Mindeststandard verfehlten (= maximal Stufe I erreicht) bzw. (2) den Regelstandard verfehlten (= maximal Stufe II erreicht). Schülerinnen und Schüler verfehlten zum späteren Messzeitpunkt die Bildungsergebnisse, wenn sie in der Hauptstudie den Mindeststandard bzw. die Noten 1 bis 3, in der Ergänzungsstudie den Regelstandard bzw. die Noten 1 bis 3 verfehlten.
Um die ermittelte Klassifikationsgüte der bildungsstandardbasierten Tests in Bezug auf die praktische Relevanz bewerten zu können, berechneten wir die Klassifikationsgüte in Hinblick auf unterschiedliche Strategien zur Identifikation „gefährdeter“ Kinder. Mit Strategie „T“ erfolgte die Klassifikation „gefährdeter“ Kinder auf der alleinigen Grundlage der Testinformation. Mit Strategie „N“ erfolgte die Klassifikation „gefährdeter“ Kinder auf der alleinigen Basis von Schulnoten. Mit Strategien „TuN“ bzw. „ToN“ erfolgte die Klassifikation „gefährdeter“ Kinder unter zwei möglichen Strategien zur Kombination der Informationen von bildungsstandardbasierten Tests und Schulnoten. Auf der Grundlage der ODER-Strategie „ToN“ werden all diejenigen Kinder als „gefährdet“ klassifiziert, die die Schwelle des Tests *oder* der Note desselben Faches unterschreiten. Mit dieser Strategie werden beispielsweise in Mathematik mit der Orientierung am Mindeststandard diejenigen Kinder als „gefährdet“ klassifiziert, die in Mathematik entweder (a) im Test den Mindeststandard verfehlen oder (b) die Note 4, 5 oder 6 erhalten sowie (c) Kinder, auf die beides zutrifft. Hingegen werden bei der UND-Strategie „TuN“ alleinig diejenigen Kinder als „gefährdet“ klassifiziert, die die Schwelle des Tests *und* der Note desselben Faches unterschreiten, zum Beispiel diejenigen Kinder, die in Mathematik sowohl im Test den Mindeststandard verfehlen und zugleich die Note 4, 5 oder 6 erhalten. Zusammenfassend zeigten sich differenzierte Befunde in der Klassifikationsgüte zur Identifizierung „gefährdeter“ Kinder auf der Grundlage bildungsstandardbasierter Tests. Insbesondere zeigten sich deutliche Unterschiede in Abhängigkeit des genutzten Schwellenwertes (Mindeststandard vs. Regelstandard) zum ersten Messzeitpunkt, sowie in Abhängigkeit der genutzten Strategie zur Identifikation der Schülerinnen und Schüler zum ersten Messzeitpunkt.

Konkret für Schülerinnen und Schüler an SMBG: Für diese lag die Sensitivität für die Tests „T“ stets über der der Noten „N“ sowohl für die Prognose im Testkriterium (Verfehlen des Mindeststandards) als auch des Notenkriteriums (Verfehlen der Noten 1, 2 oder 3). Orientierte

man sich im Test an den Regelstandards als Schwellenwert, verbesserte sich stets die Sensitivität. Wurden Kinder als „gefährdet“ klassifiziert, die die Schwelle des Tests oder der Note „ToN“ desselben Faches unterschritten, erhöhte sich stets die jeweilige Sensitivität gegenüber den anderen Strategien. Wurden Kinder als „gefährdet“ klassifiziert, wenn Sie die Schwelle der Tests und der Note „TuN“ desselben Faches unterschritten, reduzierte sich stets die Sensitivität im Vergleich zu den anderen Strategien. Die positiv prädiktiven Werte für die Noten „N“ waren höher als für die Tests „T“ in Bezug auf beide Leistungskriterien. Mit der Orientierung am Regelstandard verminderte sich der positiv prädiktive Wert der Tests. Durch „TuN“ ließ sich eine Verbesserung des positiv prädiktiven Wertes erzielen, die jedoch unterhalb der der Note „N“ blieb. Für Schülerinnen und Schüler an GY zeigte sich zusammenfassend ein vergleichbares Ergebnismuster, jedoch fielen die absoluten Werte für die Sensitivität und den positiv prädiktiven Wert, sowie die Unterschiede in Abhängigkeit der Strategien und Schwellenwerte geringer aus. Für die replizierten Analysen in der Ergänzungsstudie zeigte sich weitestgehend ein vergleichbares Ergebnismuster wie für Schülerinnen und Schüler an SMBG.

Zusammenfassend deuten die vorliegenden Ergebnisse darauf hin, dass bildungsstandardbasierte Tests – in Abhängigkeit des gewählten Schwellenwertes zum Ausgangspunkt der Prognose – vergleichbare (wenn nicht sogar bessere) Sensitivitäten und positiv prädiktive Werte aufweisen können als alternative kommerziell erhältliche, standardisierte Schulleistungstests (Studie III, EMS 1). Des Weiteren wird veranschaulicht, wie stark sich die diagnostische Entscheidung potenziell durch die Kombination von bildungsstandardbasierten Tests und Schulnoten verbessern kann. Die Wahl der Kombinationsstrategie wirkt sich unterschiedlich auf die einzelnen Indizes der Klassifikationsgüte aus. Daher sollte die Wahl der Kombinationsstrategie in Hinblick auf die diagnostische Zielstellung unter Abwägung der möglichen positiven und negativen Konsequenzen erfolgen (Studie III, Diskussion „Nutzen für die pädagogisch-psychologische Diagnostik an Schulen“).

5 Gesamtdiskussion

5.1 Zusammenfassung zentraler Befunde zur Prognose des zukünftigen Schulerfolgs

Seit 2003 und 2004 wird den Ländern in Form der Bildungsstandards für bestimmte Fächer ein bundesweit einheitlicher Referenzrahmen zur Verfügung gestellt um die Leistungen ihrer Schülerinnen und Schüler einzuordnen (KMK, 2016). Die Bildungsstandards formulieren „fachliche und fachübergreifende Basisqualifikationen, die für die weitere schulische und berufliche Ausbildung von Bedeutung sind und die anschlussfähiges Lernen ermöglichen“ (KMK, 2004, S. 7). Dieser Anspruch impliziert, dass für Schülerinnen und Schüler auf der Grundlage der Ergebnisse in bildungsstandardbasierten Tests zentrale Kriterien des Schulerfolgs vorhersagbar sein sollten. Wie gut mit bildungsstandardbasierten Tests – wie beispielsweise den Vergleichsarbeiten (VERA) – die Prognose auf den späteren Schulerfolg gelingt, wurde bisher kaum systematisch untersucht, obwohl VERA-Tests in Deutschland weit verbreitet sind: So sind Lehrkräfte der 3. und 8. Jahrgangsstufe öffentlicher Schulen in Deutschland dazu verpflichtet, jährlich in mindestens einem Fach bildungsstandardbasierte Tests – in Form der Vergleichsarbeiten (VERA) – für ihre insgesamt etwa 1.4 Millionen Schülerinnen und Schüler durchzuführen (KMK, 2012; Statistisches Bundesamt, 2017).

Aus allen drei Studien der vorliegenden Dissertationsschrift liegen auf der Grundlage von Korrelations– und Regressionsanalysen vergleichbare Befunde zur Ausprägung (1) der alleinigen Prognosegüte (s. Tabelle 5.1) und (2) des inkrementellen prognostischen Mehrwerts (s. Tabelle 5.2) bildungsstandardbasierter Tests vor. So berücksichtigen wir hier auch Werte aus den Korrelationstabellen der elektronischen Supplemente der Studie II und III, die nicht im Fokus der jeweiligen Studien standen (Ergebnisse in Studie II auf Grundlage der Berechnungen für die Gesamtpopulation der jeweiligen Schulformen in ESM 6; sowie in Studie III aus der Korrelationstabelle in ESM 4). Zudem wurden die Kennwerte für die Zusammenfassung über die Studien hinweg in vergleichbare Kennwerte transformiert und weichen infolgedessen z. T. von den berichteten Kennwerten in den jeweiligen Manuskripten ab.

5.1.1 Prognosegüte bildungsstandardbasierter Tests

Einen Überblick über die alleinige Prognosegüte bildungsstandardbasierter Tests aller Einzelstudien hinweg für das Fach Mathematik und Deutsch (Lesen) der jeweiligen

Prognosezeiträume im Vergleich zum bisher belastbarsten Forschungsstand zu VERA-Tests und kommerziell erhältlichen, standardisierten Schulleistungstests lässt sich Tabelle 3 entnehmen.

Tabelle 3

Überblick über die (alleinige) Prognosegüte (r) bildungsstandardbasierter Tests

Test	K	A	Dauer	Prognosegüte (r)		FS: VERA[a]			FS: Kommerzielle Tests[d]		
A) Mathematik				Min	Max	MD	25	75	MD	25	75
Kriterium Testleistung											
B	B	3	1	.56	.75						
B/V	B/V	6	2	.60	.70	.79	.78	.79	.68	.64	.70
B	B	1	3	.58		.68	.65	.69			
V	V	2	5	.52	.53	.64	.61	.67			
Kriterium Schulnote											
B	N	3	1	.54	.67						
B/V	N	6	2	.48	.57						
B	N	1	3	.51							
V	N	2	5	.30	.36						
B) Deutsch (Lesen)											
Kriterium Testleistung											
B	B[c]	1	1	.60							
B	B[c]	1	2	.61		.74	.70	.78	.58	.46	.68
V	V	2	5	.35	.43	.64	.63	.65			
Kriterium Schulnote											
B	N	1	1	.47							
B	N	1	2	.50							
V	N	2	5	.25	.30						

Anmerkungen. Ermittelte (alleinige) Prognosegüte innerhalb der drei Studien der vorliegenden Dissertation im Vergleich zum Forschungsstand – aufgegliedert nach der Prognosedauer. Test = Kompetenztest, K = Kriterium, A = Anzahl der Kennwerte, Dauer = Prognosedauer in Jahren, FS = Forschungsstand, VERA = Vergleichsarbeiten, Min = Minimum, Max = Maximum, MD = Median über die Korrelationskoeffizienten des Forschungsstandes, 25 = 25 %-Quantil über die Korrelationskoeffizienten des Forschungsstandes, 75 = 75 %-Quantil über die Korrelationskoeffizienten des Forschungsstandes, B = kommerzieller bildungsstandardbasierter Test/ IQB-Aufgabenpool (für Details s. Manuskripte der jeweiligen Studien), V = Vergleichsarbeiten (VERA), N = Note.

[a] C. Nachtigall, persönl. Mitteilung, 22.10.2018

[d] s. Studie I, ESM 1 bzw. Studie III, ESM 8

[c] z. T. keine expliziten bildungsstandardbasierten Aufgaben (z. B. ELEMENT-Aufgaben s. Details Studie I)

In unseren Ergebnissen für bildungsstandardbasierte Mathematiktests zeigt sich erwartungskonform die leichte Tendenz, dass die Prognosegüte mit zunehmender Prognosedauer sukzessive abnimmt. Diese systematische Abnahme der Prognosegüte mit zunehmender Prognosedauer wird auch als *validity degradation* bezeichnet (Dahlke et al., 2018). Diese leichte Tendenz in Bezug auf spätere Testleistungen zeichnet sich ebenso im

bisherigen Forschungsstand zu VERA-Daten ab (C. Nachtigall, persönl. Mitteilung, 22.10.2018). Die Prognosegüte bildungsstandardbasierter Tests auf spätere Testleistungen fällt (bis auf die Prognosedauer über 5 Jahre) weitestgehend vergleichbar mit der Prognosegüte auf spätere Schulnoten in Mathematik aus.

Unsere Ergebnisse für bildungsstandardbasierte Deutschtests im Lesen sind weniger fundiert als für die Mathematiktests. Es zeigte sich jedoch eine vergleichbare leichte Tendenz der Abnahme der Prognosegüte mit zunehmender Prognosedauer wie für die Mathematiktests. Die Prognosegüte bildungsstandardbasierter Lesetests auf spätere Lesetestleistungen fällt tendenziell etwas höher aus im Vergleich zur Prognosegüte auf spätere Schulnoten in Deutsch.

Vergleicht man unsere Befunde für bildungsstandardbasierte Mathematik- und Deutschtests für die Prognose späterer Testleistungen mit der bisher belastbarsten Befundlage dazu von Nachtigall fällt auf, dass unsere Ergebnisse für die jeweiligen Prognosezeiträume tendenziell niedriger ausfallen. Ein möglicher Grund könnte in der abweichenden Analysestrategie bei der Auswertung liegen: So hat Nachtigall (C. Nachtigall, persönl. Mitteilung, 22.10.2018) stets die Korrelationswerte auf der Grundlage der Gesamtstichprobe berechnet – ohne eine Unterteilung der Schülerinnen und Schüler nach unterschiedlichen Schulformen vorzunehmen. Unsere Befunde basieren jedoch stets lediglich auf allen Schülerinnen und Schülern einer bestimmten Schulform (der Primarstufe oder besuchten Schulform zum Zeitpunkt der Sekundarstufe). Es ist zu erwarten, dass sich die resultierenden Korrelationskoeffizienten in Abhängigkeit dieser Auswertungsstrategien unterscheiden. Die separate Auswertung nach Schulform (Strategie in unseren Studien) geht zwangsläufig mit einer Varianzeinschränkung der Schülerleistungen einher. Deshalb sollte diese mit systematisch niedrigeren Korrelationswerten im Vergleich zur Auswertung ohne Berücksichtigung der Schulform und somit ohne Varianzeinschränkung einhergehen (Sedlmeier & Renkewitz, 2013). Somit kann die gewählte Strategie in unseren Studien in der Regel als eine konservative Analysestrategie zur Bestimmung der Prognosegüte von Testverfahren bewertet werden.

Für kommerziell erhältliche, standardisierte Schulleistungstests liegt die solideste und vergleichbarste Befundlage im Vergleich zu unseren Studien für die Prognose zukünftiger Testleistungen innerhalb der Primarstufe vor. Die Ergebnisse der Sekundarstufe stellen zu strenge Vergleichswerte dar, da hier meist nur ein Monat zwischen dem ersten und zweiten

Messzeitpunkte lag (s. Abschnitt 3.1.2 „Prognosegüte zukünftiger Testleistungen“, Tabelle 1). Es zeigt sich, dass unsere Ergebnisse aus den Einzelstudien über eine Prognosedauer von 2 Jahren vergleichbar gut ausfallen wie die bisher berichteten Befunde über einen vergleichbaren Zeitraum für kommerziell erhältliche Tests.

Wie schneiden unsere Ergebnisse zu bildungsstandardbasierten Mathematik- und Deutschtests im Vergleich zu internationalen Befunden ab? Einen guten Überblick über die Prognosegüte von Screeningverfahren im Fach Mathematik liefert Gersten mit Kolleginnen und Kollegen (2012). Der Überblick beschränkt sich auf Tests, die im Kindergarten und der ersten Klassenstufe eingesetzt werden, um spätere Schwächen in der Mathematikleistung zu prognostizieren. Für die meisten Tests beziehen sich die Kennwerte auf eine Prognosedauer von maximal einem Jahr, nur selten über eine Zeitspanne von 1 oder 2 Jahren. Die Prognosegüte für diese Screeningverfahren im Fach Mathematik liegt zwischen $.46 \leq r \leq .63$ (= 25 %-Quantil ≤ 75 %-Quantil). Unsere Befunde für bildungsstandardbasierte Tests zur Prognose zukünftiger Testleistungen fallen somit vergleichbar aus. Des Weiteren liegen bisher zwei Meta-Analysen zu einem spezifischen Lesekompetenztest im englischsprachigen Raum im Zusammenhang mit anderen standardisierten Lesekompetenztests vor. In der einen Metaanalyse (Reschly, Busch, Betts, Deno & Long, 2009) erfolgte eine Zusammenfassung von insgesamt 289 Korrelationskoeffizienten für Studien innerhalb der 1. bis 6. Klassenstufe. Der mittlere Korrelationskoeffizient lag hier bei .68, zwischen $.61 \leq r \leq .74$ (= 25 %-Quantil ≤ 75 %-Quantil). Ähnliche Ergebnisse zeigten sich für die zweite diesbezügliche Meta-Analyse (Yeo, 2010). Unsere Befunde für bildungsstandardbasierte Deutschtests im Lesen zur Prognose zukünftiger Testleistungen fallen somit vergleichbar aus. Zudem zeigten sich im Rahmen der Moderatoranalysen der beiden Meta-Analysen, dass erneut mit zunehmender Dauer der Prognose die Prognosegüte abnahm und bestätigen somit ebenfalls die bereits erwähnte *validity degradation* (Dahlke et al., 2018).

Zusammenfassend wird deutlich, dass die Befundlage zu bildungsstandardbasierten Tests zur Prognose des Schulerfolgs in Form von bildungsstandardbasierten Testleistungen und Schulnoten aktuell für deutschsprachige Leistungstests in Mathematik und Deutsch (Lesen) auf der umfassendsten Datenlage im deutschsprachigen Raum für standardisierte Leistungstests basiert. Es zeigt sich ein relativ kohärentes Befundmuster zur Prognosegüte bildungsstandardbasierter Tests. Dieses gibt Hinweise für eine langfristige Prognosegüte

bildungsstandardbasierter Tests von bis zu 5 Jahren, welche in ihrer Ausprägung – sofern vergleichbare Werte des bisherigen Forschungsstandes vorliegen – mit den Befunden zu anderen standardisierten Leistungstests des nationalen und internationalen Forschungsstandes vergleichbar sind. Dies untermauert empirisch, dass bildungsstandardbasierte Tests durchaus das Potenzial zur Prognose zukünftiger schulischer Leistungen besitzen.

5.1.2 Inkrementeller prognostischer Mehrwert bildungsstandardbasierter Tests

Einen Überblick über den inkrementellen prognostischen Mehrwert bildungsstandardbasierter Tests aller Einzelstudien für das Fach Mathematik und Deutsch (Lesen) der jeweiligen Prognosezeiträume im Vergleich zur (alleinigen) Prognosegüte bildungsstandardbasierter Tests aller Einzelstudien lässt sich Tabelle 4 entnehmen.

Tabelle 4

Inkrementeller prognostischer Mehrwert (r_{adj} , sr) bildungsstandardbasierter Tests

Test	K	A	Dauer	Mehrwert (r_{adj})		Mehrwert (sr)		Prognosegüte (r)	
A) Mathematik				Min	Max	Min	Max	Min	Max
Kriterium Testleistung									
B	B	2	1	.30	.45	.24	.33	.56	.75
B/V	B/V	6	2	.36	.56	.28	.48	.60	.70
B	B	1	3	.33		.26		.58	
Kriterium Schulnote									
B	N	1[a]	1	.22		.17		.54	.67
B/V	N	6	2	.12	.24	.10	.17	.48	.57
B	N	1	3	.15		.10		.51	
B) Deutsch (Lesen)									
Kriterium Testleistung									
B	B[b]	1	1	.44		.36		.60	
B	B[b]	1	2	.43		.36		.61	
Kriterium Schulnote									
B	N	1	1	.19		.10		.47	
B	N	1	2	.19		.14		.50	

Anmerkungen. Überblick über den inkrementellen prognostischen Mehrwert in Form von adjustierten Korrelationskoeffizienten (r_{adj}) und Semipartialkorrelationskoeffizienten (sr) der drei Studien der vorliegenden Dissertation im Vergleich zur ermittelten alleinigen Prognosegüte (r) im Rahmen der drei Studien – aufgegliedert nach der Prognosedauer. Test = Kompetenztest, K = Kriterium, A = Anzahl der Kennwerte, Dauer = Prognosedauer in Jahren, Min = Minimum, Max = Maximum, B = kommerzieller bildungsstandardbasierter Test/ IQB-Aufgabenpool (für Details s. Manuskripte der jeweiligen Studien), V = Vergleichsarbeiten (VERA), N = Note.

[a] Kennwert auf Studie I ohne Kontrolle für Schulnote mit r = .41 bzw. sr = .37 nicht berücksichtigt

[b] z. T. keine expliziten bildungsstandardbasierten Aufgaben (z. B. ELEMENT-Aufgaben s. Details Studie I)

Der inkrementelle prognostische Mehrwert bildungsstandardbasierter Tests wurde durch zwei Kennwerte quantifiziert: r_{adj} bildet den inkrementellen Wert der Tests auf das Kriterium bei gleichzeitiger Kontrolle der Schulnote ab (und z. T. weiterer leistungsprädiktiver Merkmale wie sozioökonomischer Hintergrund und Intelligenz, s. Details Studie I). Hingegen informiert *sr* über den uniquen korrelativen Zusammenhang der Testleistung mit dem jeweiligen Kriterium, wenn die gemeinsame Varianz zwischen Note und dem jeweiligen Kriterium herausgerechnet wird (Cohen, Cohen, West & Aiken, 2003). Erwartungskonform sind die Kennwerte für r_{adj} höher ausgeprägt als für *sr*. Für beide Kennwerte zeigt sich nur z. T. eine Tendenz zur Abnahme der Prognosegüte mit zunehmender Prognosedauer. Der inkrementelle prognostische Mehrwert ($r_{adj,}$ *sr*) bildungsstandardbasierter Mathematik- und Deutschtests zur Prognose späterer bildungsstandardbasierter Testleistungen fällt stets höher aus im Vergleich zur Prognose späterer Schulnoten im selben Fach. Beim Vergleich des inkrementellen prognostischen Mehrwerts mit der alleinigen Prognosegüte der Tests liegen insbesondere die

Wertebereiche für die Mathematiktests zur Prognose der Testleistung von r_{adj} und r am stärksten beieinander.

Die in Tabelle 4 zusammenfassend dargestellten Befunde zum inkrementellen prognostischen Mehrwert bildungsstandardbasierter Tests basieren auf unterschiedlichen Modellen, die sich in der Anzahl der zusätzliche berücksichtigten leistungsprädiktiven Merkmale unterscheiden (v. a. Studie I). Hier wurden jedoch stets die Kennwerte für Modelle zusammengefasst, die mindestens die Schulnote als weiteres leitungsprädiktives Merkmal berücksichtigten. In den Modellen war die Schulnote z. T. ein sehr strenger Vergleichsmaßstab, da diese z. T. zu einem späteren Zeitpunkt als zum Zeitpunkt des Testergebnisses herangezogen wurde, wodurch die Prognosegüte der Schulnote aufgrund der *validity degredation* tendenziell höher ausfallen sollte und somit systematisch begünstigt ist. Folglich sind unsere zusammengefassten Befunde zum inkrementellen prognostischen Mehrwert – bei denen zumindest immer die Schulnote berücksichtigt wurde – eher als eine Untergrenze für die Abschätzung der praktischen Relevanz bildungsstandardbasierter Tests zu werten.

Der deutschsprachige Forschungsstand zum inkrementellen prognostischen Mehrwert von standardisierten kommerziell erhältlichen und nicht kommerziell erhältlichen Leistungstests ist sehr überschaubar und die wenigen Studien variieren zudem in zahlreichen Studienparametern (wie bspw. den zusätzlichen leistungsprädiktiven Merkmalen). Zudem sind die Analysemodelle der Studien weniger streng als in unseren, sodass hier hinreichende Befunde zur Einordnung unserer Ergebnisse fehlen. Die hiermit vorliegende umfangreiche empirische Befundlage zum inkrementellen prognostischen Mehrwert bildungsstandardbasierter Tests liegt bisher unseres Wissens in diesem Ausmaß für keinen anderen deutschsprachigen, standardisierten Schulleistungstest vor.

Auch wenn die Ausprägung der Höhe der Korrelationskoeffizienten hier nicht abschließend bewertet werden kann, so liefern diese empirische Hinweise, dass bildungsstandardbasierte Tests einen inkrementellen prognostischen Mehrwert gegenüber Halbjahresnoten haben können. Und dies, obwohl die Testergebnisse lediglich auf einer einmaligen Messung basieren und die berücksichtigten Halbjahresnoten, die sich aus mehreren Leistungsmessungen zusammensetzen, reliabler sein sollten.

5.1.3 Einzelschulen: Generalisierbarkeit der Prognosegüte

Die zusammengefassten Ergebnisse unter 5.1.1 zur (1) alleinigen Prognosegüte und unter 5.1.2 zum (2) inkrementellen prognostischen Mehrwert bildungsstandardbasierter Tests wurden stets auf der Grundlage der Leistungsergebnisse im Test aller Schülerinnen und Schüler der Gesamtstichprobe (einer jeweiligen Schulform) ermittelt. Dies ist ein typisches Vorgehen in den meisten Studien zur Untersuchung der psychometrischen Güte von Testverfahren. Es wird nicht hinterfragt, inwiefern die Prognosegüte, die für die Gesamtstichprobe ermittelt wurde, sich auf die Schülerschaft an den Einzelschulen generalisieren lässt. Oder anders formuliert: Inwiefern es Unterschiede zwischen Schulen in der Prognosegüte bildungsstandardbasierter Tests gibt.

Die mittleren (aggregierten) Kennwerte der Prognosegüte über eine Prognosedauer von 2 Jahren über die verschiedenen Einzelschulen von SMBG (Kriterium Test: $\beta = .66$, $\beta_{adj} = .48$, $sr = .39$; Kriterium Note: $\beta = .57$, $\beta_{adj} = .24$, $sr = .16$) bzw. GY (Kriterium Test: $\beta = .60$, $\beta_{adj} = .43$, $sr = .33$; Kriterium Note: $\beta = .56$, $\beta_{adj} = .22$, $sr = .14$) fielen in ihrer Ausprägung vergleichbar mit dem bisherigen Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen standardisierten Mathematiktests aus.

Die Befunde zur Heterogenität der Güte in Bezug auf spätere bildungsstandardbasierte Testleistungen und Schulnoten zeigen, dass sich (1) die alleinige Prognosegüte und (2) der inkrementelle prognostische Mehrwert der bildungsstandardbasierten Tests gegenüber Schulnoten z. T. nur eingeschränkt zwischen Einzelschulen von SMBG und Einzelschulen von GY generalisieren lässt. Jedoch liefert unsere Studie ebenso erstmalige empirische Evidenz dafür, dass VERA-Mathematiktests spätere Testleistungen und Schulnoten an der Mehrzahl der Schulen einer jeweiligen Schulform prognostizieren können. Zudem weisen die Befunde darauf hin, dass die Tests an den meisten Schulen einen (z. T. deutlichen) prognostischen Mehrwert für die Leistungsdiagnostik neben Zeugnisnoten haben können.

Die höchste Heterogenität in der Güte des Tests zeigt sich eher zwischen SMBG in Bezug auf die Prognosegüte auf spätere Schulnoten und den inkrementellen prognostischen Mehrwert in Bezug auf spätere Testleistungen und Schulnoten. Unterschiede in Bezug auf den prognostischen Mehrwert zwischen den SMBG geben Hinweise, dass an manchen dieser Schulen die Halbjahresnoten mehr oder weniger redundante Informationen mit bildungsstandardbasierten Tests aufweisen. So werden an manchen SMBG sehr ähnliche

Informationen im Vergleich zu den bildungsstandardbasierten Tests erfasst, sodass der Mehrgewinn gegenüber den Noten niedrig ist. Jedoch gibt es auch einen erheblichen Teil an SMBG, an denen die Schulnoten kaum redundante Leistungsinformationen mit den Tests teilen. D. h. der Mehrgewinn gegenüber den Noten ist an diesen SMBG sehr hoch ist und somit scheint v. a. an diesen Schulen viel Potenzial zur Verbesserung der Leistungsdiagnostik zu bestehen. Zwischen GY gab es deutlich weniger Heterogenität in der Güte der Tests, die sich v. a. auf die Güte der Tests in Bezug auf spätere Testleistungen beschränkte. Die meisten Kennwerte zum inkrementellen prognostischen Mehrwert weisen keine statistisch signifikante Heterogenität auf.

Abschließend wurde aus den Moderatoranalysen deutlich, dass sich die vorliegende Heterogenität zwischen den Einzelschulen der jeweiligen Schulform in der Prognosegüte und dem inkrementellen Mehrwert z. T. durch schulspezifische Merkmale erklären lässt. Für praxisrelevante Implikationen sind hier v. a. die Befunde zum schulspezifischen Leistungsniveau relevant. So zeigte sich, dass sich Einschränkungen in der Validitätsgeneralisierung über Einzelschulen zwischen GY durch eine Anpassung der Testhefte an das Leistungsniveau der Schülerschaft z. T. vollständig reduzieren lassen und unterfüttert somit empirisch aktuelle Bestrebungen der KMK in Bezug auf die Weiterentwicklung von VERA (KMK, 2012). So wird durch eine Flexibilisierung und Modularisierung der VERA-Testdurchführung angestrebt, passgenauere Testmaterialien (bspw. zum Leistungsniveau der Schülerinnen und Schüler) an den Schulen zu realisieren. Potenziell sollte sich dies auch an SMBG positiv auf die Prognosegüte auswirken, jedoch konnten diesbezügliche Effekte im Rahmen unserer Studie vermutlich aufgrund von Methodenartefakten nicht adäquat abgebildet werden (s. Studie III, ESM 4).

5.1.4 Screeningfunktion: Identifikation „gefährdeter" Kinder

Aus wissenschaftlicher Sicht wird das Potenzial bildungsstandardbasierter Test für die individualdiagnostische Anwendung z. T. darin gesehen, dass diese im Sinne eines Screenings Hinweise für die weiterführende Diagnostik geben können (Köller & Reiss, 2013; Leutner et al., 2008). Die bisher berichteten Befunde für bildungsstandardbasierte Tests für die Prognosegüte auf der Basis von Korrelations- und Regressionsanalysen sind nicht hinreichend um die Qualität dieser potenziellen Screeningfunktion zu beurteilen. Entscheidend ist hier die Analyse der Klassifikationsgüte, die wir mit etablierten Indizes, wie der Sensitivität und dem positiv prädiktiven Wert, bestimmten.

Aus den Ergebnissen wurde deutlich, dass bildungsstandardbasierte Tests z. T. vergleichbare (wenn nicht sogar bessere) Sensitivitäten und positiv prädiktive Werte aufweisen können als alternative kommerziell erhältliche, standardisierte Schulleistungstests. Dies ist unter anderem abhängig davon, welcher Schwellenwert in bildungsstandardbasierten Tests (Mindeststandard vs. Regelstandard) zum Ausgangspunkt der Prognose für die Klassifikation der Schülerin bzw. des Schülers als „gefährdet“ herangezogen wird. Wählt man die weniger kritische Schwelle (Regelstandard), so werden mehr „gefährdete“ Schülerinnen und Schüler richtigerweise identifiziert. Aber dies geht gleichzeitig auch mit einer höheren Anzahl an Schülerinnen und Schülern einher, die fälschlicherweise als „gefährdet“ klassifiziert wurden. Für die Wahl des relevanten Schwellenwertes sollte eine Abwägung der Konsequenzen in Hinblick auf die diagnostische Zielstellung erfolgen (s. Diskussion der Studie III zum Abschnitt „Nutzen für die pädagogisch-psychologische Diagnostik an Schulen“). Hier wird jedoch u. a. deutlich, dass die Einordnung der Leistungen von Schülerinnen und Schülern anhand der fünf Kompetenzstufen hilfreich zur Unterstützung der Diagnostik sein kann, da hier potenziell die Leistungen der Schülerinnen und Schüler aufgrund mehrerer Schwellenwerten eingeordnet werden können. Wichtig ist mir hierbei jedoch zu betonen, dass diese Argumentation kein Plädoyer für eine differenziertere Klassifikation der Leistungen von Schülerinnen und Schüler in Bezug auf die Kompetenzstufen sein soll. Die verschiedenen Schwellenwerte bieten lediglich eine flexiblere Diagnostik an den Schulen in Abhängigkeit des Leistungsniveaus der Schülerschaft (Basis- und Selektionsraten) und der bereits benannten diagnostischen Zielstellungen (bspw. möglichst viele „gefährdete“ Schülerinnen und Schüler rechtzeitig identifizieren). Diese Zielstellung sollte vor dem Hintergrund einer Screeningfunktion der Tests verfolgt werden – also stets mit dem Ziel einer dichotomen Klassifikation der Leistungen der Schülerinnen und Schüler in „gefährdete“ und „nicht gefährdete“ vorzunehmen. Diese dichotome Klassifikation setzt grundlegend einen geringeren Anspruch an die Reliabilität des Verfahrens im Vergleich zur Klassifikation in weiterführenden Stufen. Mit jeder weiteren Klassifikationsstufe steigt die Anzahl an Fehlklassifikationen (Ercikan & Julian, 2002). Somit steigt der Anspruch an die Reliabilität des Testverfahrens, welche nach bisherigen ersten empirischen Hinweisen für bildungsstandardbasierte Tests im Rahmen von VERA nicht hinreichend zu bestehen scheint (Pant et al., 2017; Tiffin-Richards, 2011).

Darüber hinaus wird im Rahmen der Befunde zur Klassifikationsgüte veranschaulicht, dass sich die diagnostische Entscheidung potenziell durch die Kombination von

bildungsstandardbasierten Tests und Schulnoten verbessern kann. Die Wahl der Kombinationsstrategie wirkt sich unterschiedlich auf die einzelnen Indizes der Klassifikationsgüte aus. Daher sollte auch hier die Wahl der Kombinationsstrategie in Hinblick auf die diagnostische Zielstellung unter Abwägung der möglichen positiven und negativen Konsequenzen erfolgen. Vor allem wird anhand der Ergebnisse verdeutlicht, dass auch die Übereinstimmung von Schulnoten und Tests in Bezug auf die Klassifikation „gefährdeter" Kinder die Güte der Klassifikation verbessern kann und somit diese bestätigende Information zur „Absicherung" der Diagnose (Strategie TuN) zu einer Erhöhung des positiv prädiktiven Wertes beiträgt. Dies wird jedoch anscheinend von Lehrkräften im Rahmen von VERA nicht als zusätzliche relevante Information erkannt und somit nicht in diagnostischen Prozessen berücksichtigt (Ramsteck & Maier, 2015).

5.2 Diskussion im Rahmen formativer Leistungsmessung an Schulen

5.2.1 Einordnung der Studienbefunde

Formative Leistungsmessung gilt als eines der wirksamsten Rahmenkonzepte zur Förderung schulischen Lernens. Man versteht darunter die „lernprozessbegleitende Beurteilung von Leistungen mit dem Ziel, diese diagnostischen Informationen zu nutzen, um Unterricht und letztlich das individuelle Lernen zu verbessern" (Schütze et al., 2018, S. 698). Insbesondere wird von Black und Wiliam (2009) betont, dass Leistungsdaten im Rahmen der formativen Leistungsmessung unterstützen sollen, Entscheidungen im weiteren Lehr-Lern-Prozess zu treffen. Diese Leistungsdaten können auf der Grundlage unterschiedlichster Messinstrumente bzw. Testverfahren gewonnen werden (Köller, 2005). Dem Testverfahren an sich liegt es jedoch nicht inne ob es formativ ist oder nicht. Es ist die Nutzung des Testverfahrens, das formativ erfolgen kann. Als Gegenpool zur formativen Nutzung wird meist von summativer Nutzung gesprochen, d. h. im Rahmen der Leistungsfeststellung das Ziel verfolgt, Leistungen zusammenfassend zu beurteilen. Die Leistung wird somit häufig nach Abschluss einer Maßnahme erfasst (Schütze et al., 2018). Entsprechend kann die Nutzung von Testverfahren mehr oder weniger stark auf einem Kontinuum zwischen diesen beiden Nutzungsformen variieren.

Die jährlich bundesweit in Deutschland zum Zeitpunkt der 3. und 8. Jahrgangsstufe an allen öffentlichen Schulen durchgeführten bildungsstandardbasierten Tests im Rahmen der Vergleichsarbeiten (VERA) werden grundlegend dem Konzept der formativen

Leistungsmessung zugeordnet, selbst wenn konstatiert wird, dass der summative Charakter bei der Umsetzung der VERA-Tests überwiegt (Maier, 2010; Schütze et al., 2018). Auch wenn die VERA-Tests nicht alle Merkmale einer formativen Nutzung erfüllen, so weisen die Tests ein zentrales Schlüsselmerkmal formativer Leistungsmessung auf – eine kriteriale Bezugsnorm (Maier, 2010; Schütze et al., 2018). So beschreiben die Bildungsstandards „die fachbezogenen Kompetenzen einschließlich zugrunde liegender Wissensbestände, die Schülerinnen und Schüler bis zu einem bestimmten Zeitpunkt ihres Bildungsganges erreicht haben sollen“ (Kultusministerkonferenz, 2006, S. 6). Folglich werden konkrete Lernziele und Erfolgskriterien kommuniziert. Potenziell könn(t)en Lehrkräfte – als ein weiteres wichtiges Schlüsselmerkmal formativer Leistungsmessung – mit Hilfe der VERA-Tests den individuellen Lernstand der Schülerinnen und Schüler in Bezug zu diesem Lernziel erfassen, bzw. den Leistungstand auf der Bildungsstandardmetrik zurückgemeldet bekommen. Eine Rückmeldung auf der kriterialen Metrik der Bildungsstandards erhalten die Lehrkräfte jedoch nur in einem Teil der Bundesländer (Isaac, 2013; Pant et al., 2017).

Warum erfolgt die Rückmeldung über Leistungsstände nicht bundesweit auf der kriterialen Bildungsstandardmetrik im Rahmen der VERA-Tests? Aus wissenschaftlicher Sicht ermöglicht die mangelnde Reliabilität keine hinreichend valide Einordnung der Leistung der Schülerinnen und Schüler auf Individualebene (Pant et al., 2017). Diesem wird u. a. Rechnung getragen, indem das Ungenauigkeitsintervall des Testwertes für die Schülerin bzw. den Schüler ebenfalls zurückgemeldet wird (Pant et al., 2017). Ergänzend hierzu wurde ebenso aus wissenschaftlicher Sicht eingeräumt, dass eine Verortung der Leistung der Schülerinnen und Schüler im Sinne eines Screenings denkbar wäre (Köller & Reiss, 2013; Leutner et al., 2008). Im Rahmen der Studie III haben wir daher konkret empirisch geprüft – auf der Rationale, die z. T. mehr oder weniger explizit in den Kompetenzstufenmodellen formuliert wird – inwiefern VERA-Tests in Mathematik und Deutsch (Lesen) im Sinne eines Screenings für eine Identifikation von Schülerinnen und Schülern geeignet sind, die gefährdet sind, spätere Bildungsergebnisse zu verfehlen. Auf diese Weise könnten die VERA-Tests beispielsweise Hinweise für möglichen Förderbedarf bei Schülerinnen und Schülern liefern, welche jedoch im Sinne eines Screenings mit weiteren diagnostischen Verfahren konkretisiert und abgeklärt werden sollten. Die Befunde unserer Studie geben erste empirische Hinweise, dass die Identifizierung von „gefährdeten“ Kindern auf der Grundlage von VERA-Tests im Sinne eines Screenings sogar für Leistungsprognosen von bis zu 5 Jahren hinreichend gut gelingt. Zudem zeigte sich, dass die formulierten Bildungsstandards in Form der

Kompetenzstufen den Schulen die Möglichkeit geben, unterschiedliche Schwellenwerte für das Screening zu nutzen. Damit haben Lehrkräfte an den Schulen die Möglichkeit, unter Berücksichtigung des bestehenden Leistungsniveaus der Schülerschaft an ihrer Schule die diagnostische Entscheidung auf Grundlage der VERA-Tests in Bezug auf ihre diagnostische Zielstellung hin zu verbessern. Diese spiegelt das mögliche diagnostische Potenzial der Tests an den Schulen wider.

Dass das diagnostische Potenzial von VERA-Tests durchaus zwischen Einzelschulen variieren kann, zeigten unsere Befunde aus der Studie II im Fach Mathematik – jedoch auf der Basis von Regressionskoeffizienten. Die Befunde untermauern, dass sich durch eine bessere Passung des Tests zum Leistungsniveau der Schülerschaft das diagnostische Potenzial der Tests – durch eine reliablere Leistungsmessung - steigern lässt. Bildungsstandardbasierte Tests wie VERA sind für Schülerinnen und Schüler im mittleren bis niedrigeren Leistungsniveau optimiert und somit weniger reliabel für (sehr) leistungsstarke Schülerinnen und Schüler. So ist im Kontext von raschskalierten Tests – zu denen die bildungsstandardbasierten Tests gehören – gut dokumentiert, dass die Reliabilität zunimmt, je besser die Itemschwierigkeiten des Tests zum Leistungsniveau der Personen passen (Furr, 2018). Zumindest unterstützen diese Befunde aktuelle Bestrebungen der KMK, die eine Flexibilisierung und Modularisierung der Testdurchführung im Rahmen von VERA anzielt, um durch passgenauere Testmaterialien (bspw. dem Leistungsniveau der Schülerinnen und Schüler) den diagnostischen Erkenntnisgewinn an den Schulen zu fördern (KMK, 2012). Dies könnte ein Ansatz sein, um das diagnostische Potenzial der Tests durch Erhöhung der Prognosegüte zu verbessern und somit das Potenzial der Tests zur formativen Leistungsmessung an den Schulen zu steigern. Jedoch können wir aufgrund fehlender vergleichbarer Studien nicht abschätzen, inwiefern dieser Ansatz zu einer praxisrelevanten Verbesserung des diagnostischen Potenzials führt. Trotz dessen ist sicherlich nicht zu erwarten, dass sich dadurch das diagnostische Potenzial von VERA-Tests auf Individualebene substanziell verbessern wird, sodass es annähernd der Güte einer Lernverlaufsdiagnostik im Rahmen formativer Leistungsmessung nahe kommen würde. Folglich ist das Potenzial der VERA-Tests für die Individualdiagnostik nur sehr eingeschränkt im Sinne einer formativen Leistungsmessung nutzbar. So könnten diese, nützliche Hinweise für die weiterführende Diagnostik liefern, aber weniger gute zur Ableitung direkter instruktionaler Entscheidungen in Bezug auf einzelne Schülerinnen und Schüler.

Im Rahmen der aktuellen Fassung zur „Weiterentwicklung der Vergleichsarbeiten (VERA)“ werden zudem weitere Flexibilisierungen in der Testdurchführung angestrebt, wie beispielsweise die Variation von Testterminen innerhalb eines Testzeitraumes, sowie die Auswahl getesteter Fächer bzw. Kompetenzbereiche. Dieser passgenauere Einsatz der VERA-Tests soll die Unterrichts- und Schulentwicklung vor Ort unterstützen. Somit wird zukünftig eine Weiterentwicklung von VERA angestrebt, die die formative Leistungsmessung bzw. Nutzung von VERA fördert, indem Schulen die VERA-Tests an ihre Zielstellungen in der Unterrichts- und Schulentwicklung anpassen können (u. a. Maier, 2010).

5.2.2 Bewertung des Studiendesigns: Validierung der Prognosegüte

In Abschnitt 2.1 habe ich auf der Grundlage eines argumentbasierten Ansatzes (u. a. Kane, 2013) begründet, weshalb es relevant ist, die Prognosegüte bildungsstandardbasierter Testverfahren, wie beispielsweise VERA zu untersuchen. Nun möchte ich das gewählte Studiendesign und die gewählten Analysestrategien vor dem Hintergrund dieses Ansatzes reflektieren, um die Güte der drei Validierungsstudien im Rahmen der vorliegenden Dissertation zu bewerten, v. a. mit Blick auf Implikationen für formative Leistungsmessung.

In allen drei Studien haben wir die Güte bildungsstandardbasierter Tests für die Prognose auf spätere bildungsstandardbasierte Testleistungen und Schulnoten desselben Faches untersucht. Diese stellen wichtige zukünftige Leistungskriterien dar, wenn man die Abschätzung der Prognosegüte von Testverfahren in Bezug auf formative Leistungsmessung validieren möchte. Beides sind relevante Leistungsmaße, die die schulische Leistungsentwicklung der Schülerinnen und Schüler abbilden. Wäre man verstärkt an einer Validierung mit Blick auf eine summative Leistungsmessung interessiert, wäre sicherlich relevanter gewesen, spätere schulische und vor allem berufliche Leistungskriterien im Zusammenhang mit Abschlüssen zu untersuchen wie beispielsweise Klassenwiederholungen, schulische Bildungsabschlüsse oder Prüfungs- und Abschlusszeugnisnoten.

Wie gut die Schülerinnen und Schüler in zukünftigen bildungsstandardbasierten Tests abschneiden, ist unmittelbar relevant für formative Leistungsmessung und stellt im weitesten Sinne eine Form der kriterialen Lernverlaufsdiagnostik zur Begleitung des Lernprozesses in Hinblick auf das identische Lernziel bzw. Erfolgskriterium dar – auch wenn typischerweise im Kontext formativer Leistungsmessung die Leistungsentwicklung über deutlich kürzere Zeitintervalle erfolgt (Schütze et al., 2018). Da Prädiktor und Kriterium auf dem gleichen

kriterialen Bewertungsmaßstab basieren – den Bildungsstandards –, waren in Bezug auf dieses Leistungskriterium die höchsten Ausprägungen der Prognosegüte bildungsstandardbasierter Tests zu erwarten. Im Rahmen der Studie III, zur Messung der klassifikatorischen Prognosegüte bildungsstandardbasierter Tests auf das Verfehlen von Bildungsergebnissen in Bezug auf spätere Testleistungen, haben wir besonders kritische Leistungsstände definiert, die von einer breiten Leserschaft als relevant erachtet werden sollte. Mit der Relevanz der Befunde, steigt auch der Wert der Validierung (Kane, 2017). So hatten im Rahmen unserer Ergänzungsstudie in Studie III, Kinder das Bildungsergebnis in der 6. Klassenstufe verfehlt, wenn sie den Regelstandard im Test – definiert für die 4. Klassenstufe – verfehlt hatten. Folglich zum Zeitpunkt der 6. Klassenstufe die Schülerinnen und Schüler die durchschnittlichen Leistungserwartungen verfehlt haben, die man bereits 2 Jahre früher von ihnen erwartet hätte.

Neben dem Testkriterium haben wir die Prognosegüte bildungsstandardbasierter Tests in Bezug auf spätere Schulnoten bzw. Zeugnisnoten validiert. Zeugnisnoten sind ein wichtiges etabliertes Leistungsmaß im Schulkontext. Diese setzen sich aus zahlreichen Leistungsmessungen im Schulkontext innerhalb eins Faches zusammen und weisen potenziell eine größere Nähe zum Curriculum auf als bildungsstandardbasierte Tests (Pant et al., 2011). Empirische Evidenz zur Prognosegüte bildungsstandardbasierter Tests auf spätere Schulnoten dürfte vor allem für Lehrkräfte ein wichtiges – wenn nicht sogar das entscheidende – Kriterium zur Validierung der Prognosegüte darstellen. So geben Lehrkräfte als Gründe für die Nicht-Nutzung der VERA-Daten an, dass diese nach ihrer Wahrnehmung lediglich Randthemen und keine zentralen Themen des Lehrplans fokussieren (Ramsteck & Maier, 2015).

Formative Leistungsmessung soll lernbegleitend sein und erfordert somit die fortlaufende Integration von Leistungsinformationen verschiedener Quellen, um diagnostische Entscheidungen abzuleiten. Im Rahmen der drei Studien haben wir versucht, diesen Umstand in unsere Validierung einzubeziehen. So haben wir den prognostischen Mehrwert bildungsstandardbasierter Tests unter Berücksichtigung weiterer leistungsrelevanter Merkmale im Schulkontext berücksichtigt. Dabei haben wir stets als weiteres leistungsrelevantes Merkmal die Schul(halbjahres)noten berücksichtigt, die in den meisten Schulen als weiteres Leistungskriterium zur Verfügung stehen sollten und somit die Praxisrelevanz unserer Befunde erhöht. Zudem haben wir meist eine konservative

Analysestrategie gewählt, bei welcher wir sehr strenge zusätzliche Prädiktoren gegenüber den bildungsstandardbasierten Tests einbezogen haben, die in Bezug auf die Prognosegüte (bspw. durch eine kürzere Prognosedauer) bevorteilt waren (s. Abschnitt 5.1.2 Inkrementeller prognostischer Mehrwert). Wir haben hier eine strenge Prüfung vorgenommen, um den prognostischen Mehrwert der Tests nicht zu überschätzen. Insbesondere haben wir im Rahmen der Studie III, bei der Bestimmung der klassifikatorischen Prognosegüte, deutlicher das ergänzende Potenzial von Schulnoten und bildungsstandardbasierten Tests zur Ableitung diagnostischer Entscheidungen herausgearbeitet. Hier konnten der Mehrwert jedes Leistungsindikators und der Einfluss möglicher Strategien zur Kombination beider Leistungsindikatoren für die Güte der diagnostischen Entscheidung deutlich werden. Auch wenn wir in unseren Studien – im Vergleich zur Praxis – nur in einem relativ überschaubaren Rahmen die Nutzung verschiedener Leistungskriterien zur Diagnostik untersucht haben, sind dies wichtige empirische Befunde, die ein bisher bestehendes Forschungsdesiderat in der Validierung standardisierte Leistungstests füllen.

Vor allem Black und Wiliam (2009) betonen die entscheidungsunterstützende Funktion von formativer Leistungsmessung. Inwiefern bildungsstandardbasierte Tests das Potenzial zur Verbesserung von diagnostischen Entscheidungen besitzen, haben wir im Rahmen der Studie III fokussiert, indem wir die klassifikatorische Prognosegüte untersucht haben. Auch Kane (2013) betont in seinen Erläuterungen im Rahmen seines argumentbasierten Ansatzes zur Validierung von Leistungstests, dass hierbei stets positive und negative Konsequenzen Berücksichtigung finden müssen. Dies haben wir in der Diskussion der Studie III einfließen lassen (s. Abschnitt „Nutzen für die pädagogisch-psychologische Diagnostik an Schulen“).

Zusammenfassend haben wir im Rahmen der drei Studien der vorliegenden Dissertation in Hinblick auf formative Leistungsmessung und im Sinne von Kane (2013) sukzessive die Ausprägung der Prognosegüte bildungsstandardbasierter Tests und deren Grenzen auf der Grundlage der empirischen Befunde untersucht. Dabei haben wir stets eine kritische Stellung eingenommen. So haben wir zum einen konservative Analysestrategien bevorzugt und zum anderen die „schwächste“ Testwertinterpretation in Bezug auf die Prognosegüte – die Güte auf Individualebene – untersucht. Letztlich untermauert die starke Kohärenz der Befunde zur Prognosegüte über unsere drei Studien hinweg, dass auf der Grundlage bildungsstandardbasierter Testergebnisse späterer Schulerfolg prognostiziert werden kann. Somit scheinen die Bildungsstandards auf der Grundlage des bisherigen Forschungsstandes

den Anspruch zu erfüllen, Leistungen zu definieren, die für anschlussfähiges Lernen von Bedeutung sind. Infolgedessen können sie wichtige Schlüsselmerkmale einer formativen Leistungsmessung erfüllen.

5.3 Limitationen und Ausblick

5.3.1 Messinstrumente

Die umfassendste Datengrundlage über die drei Studien zur Bestimmung der Prognosegüte bildungsstandardbasierter Tests basierte auf Daten aus regulären VERA-Testungen innerhalb eines Bundeslandes (Studie II: N = 11 054; Studie III: N = 10 939). Jedes Schuljahr werden im Rahmen von VERA neu zusammengestellte Testhefte eingesetzt. Dabei gibt es keine Überlappung der eingesetzten Items zwischen den VERA-Tests verschiedener Schuljahre. Insbesondere kommt hinzu, dass in VERA-3-Mathematiktests lediglich zwei (der insgesamt fünf) inhaltsbezogenen Leitideen geprüft werden, die zwischen den Schuljahren wechselnd eingesetzt werden (Institut zur Qualitätsentwicklung im Bildungswesen, n. d.). In zukünftigen Studien ist empirisch zu prüfen, inwiefern sich die vorliegenden Befunde, die in Mathematik und in Deutsch (Lesen) jeweils auf einer bestimmten Itemauswahl basieren, auf andere Itemkombinationen (z. B. Items zur Messung weiterer Leitideen) generalisieren lassen. Die zusammenfassend berichteten Befunde von Nachtigall (C. Nachtigall, persönl. Mitteilung, 22.10.2018) zur Prognosegüte von VERA-Tests auf spätere VERA-Testleistungen auf der Grundlage von Korrelationskoeffizienten liefern hier empirische Hinweise auf die Generalisierbarkeit. So schwankt der Interquartilsabstand innerhalb der jeweiligen Prognosezeiträume und Fächer über die neun sukzessiven Schülerkohorten bzw. VERA-Durchgänge lediglich zwischen 0.01 bis 0.08. Diese sind relativ schmal und liegen sogar im Fach Deutsch vor. In Deutsch wurden ebenso die Leistungsergebnisse auf der Grundlage der vollständigen VERA-Testhefte miteinander korreliert, welche jedoch lediglich in der Leistungsdomäne Lesen übereinstimmen und sich meist in der zweiten geprüften Leistungsdomäne unterscheiden (bspw. Schreiben, Zuhören). Aufgrund dieser Ergebnisse lässt sich vermuten, dass sich die ermittelte Prognosegüte im Rahmen unserer Studien auch auf die Prognosegüte anderer VERA-Testhefte generalisieren lassen sollte.

Im Rahmen der Studie I und Studie II (Ergänzungsstudie) wurden zwar bildungsstandardbasierte Tests bzw. bildungsstandardbasierte Testaufgaben eingesetzt (Ausnahme siehe Diskussion der Studie I für Deutsch-Testhefte zu den prognostizierten Testkriterien), aber keine expliziten VERA-Testhefte. Trotz dessen waren die Koeffizienten

weitestgehend vergleichbar mit unseren Ergebnissen, die wir auf der Grundlage der VERA-Tests ermittelt haben.

5.3.2 Testdurchführung

Im Rahmen der regulären VERA-Testungen werden die Tests in der Regel von Lehrkräften auf der Grundlage eines Manuals durchgeführt und kodiert, sowie von diesen die Daten der einzelnen Schülerinnen und Schüler eingegeben. Dies galt somit auch für die größte Datenbasis dieser Dissertation, welche im Rahmen der regulären VERA-Tests generiert wurde. Leistungsunterschiede bei VERA-8-Tests können (müssen aber nicht zwangsläufig) aus der Durchführung und Auswertung durch geschulte Testleitungen einerseits und durch Lehrkräfte andererseits resultieren (Graf, Emmrich, Harych & Brunner, 2013; Spoden, Fleischer & Leutner, 2014). Eine geringere Objektivität, die bei der Durchführung und Auswertung der VERA-Tests durch Lehrkräfte resultieren kann, würde die Reliabilität und damit auch die Prognosegüte der Testergebnisse abschwächen**.** Da in den überwiegend untersuchten Datensätzen Lehrkräfte die Durchführung und Auswertung übernahmen, sind die vorliegenden Ergebnisse – für den Hauptteil der Befunde – eher als eine Untergrenze für die Prognosegüte der Tests zu werten. Für die Hauptbefunde der Studie I und die Befunde aus der Ergänzungsstudie in Studie III trifft diese Einschränkung nicht zu, da hier geschulte Testleitungen die Tests durchführten und geschulte Kodiererinnen und Kodierer die Auswertung und Eingabe vornahmen. Im Rahmen der Zusatzanalysen in Studie I zeigten sich allenfalls kleine Unterschiede in der Prognosegüte zwischen Tests, die von Lehrkräften oder Testleitungen durchgeführt und ausgewertet wurden (Studie I: Abschnitt „Zusatzanalysen“; sowie ESM 9, 10, 11, und 12).

5.3.3 Stichproben

Unsere Befunde basierten auf Längsschnittdatensätzen aus insgesamt zwei Bundesländern[2]. Zukünftige Studien könnten Aufschluss darüber geben, inwiefern sich unsere Ergebnisse auf andere Bundesländer übertragen lassen. So können sich die Bundesländer in der Implementation der VERA-Tests unterscheiden, was sich auf die Prognosegüte auswirken könnte. Aufgrund der Erfahrungen im angloamerikanischen Raum, in denen viele Kompetenzmessungen im Schulbereich *high-stakes*-Charakter haben (Amrein-Beardsley, Berliner & Rideau, 2010) ist zu erwarten, dass Lehrkräfte den ihnen zur Verfügung stehenden

[2] Zur Vermeidung neuartiger Vergleiche werden die Bundesländer nicht genannt.

Spielraum zur Beeinflussung der Ergebnisse durch intendierte und nicht intendierte Handlungen umso mehr nutzen, je stärker sie VERA als ein Kontrollinstrument (bspw. der Schuladministration) wahrnehmen und sie für die VERA-Ergebnisse ihrer Schülerinnen und Schüler Rechenschaft ablegen müssen. Dies könnte die Durchführungs- und Auswertungsqualität durch die Lehrkräfte und somit auch die Prognosegüte beeinflussen.

Darüber hinaus mussten wir in den einzelnen Studien feststellen, dass unsere Analysen zum Teil auf positiv selektierten Stichproben, in Bezug auf die Leistungsmerkmale Testleistung und Schulnote, basierten. Dies ist zudem grundlegend für die genutzten Archivdatensätze zu den VERA-Tests zu erwarten. Diese weisen keinen systematischen Stichprobenausfall auf, der üblicherweise in Längsschnittstudien entsteht, in denen der Datensatz ausgehend von der ersten Erhebung prospektiv erstellt wird. Die Datensätze wurden jedoch ausgehend vom letzten Messzeitpunkt aneinandergefügt. Somit sind systematisch bestimmte Schülerinnen und Schüler, für welche tendenziell niedrigere Leistungen zu erwarten sind, nicht in den Längsschnitttatendsätzen enthalten. Dazu gehören beispielsweise Schülerinnen und Schüler die eine Klasse wiederholten oder die Schule wechselten. Die Varianz der Leistungen ist aufgrund dessen in unseren Datensätzen eingeschränkt und sollte in der Regel mit einer geringeren Prognosegüte einhergehen (wie sich diese auch in den Moderatoranalysen der Studie II für den Zusammenhang der schulspezifischen Leistungsheterogenität mit der Prognosegüte zeigt; sowie Sedlmeier & Renkewitz, 2013). Folglich spiegeln unsere Befunde aufgrund der untersuchten Datenbasis eher die Untergrenze der Prognosegüte bildungsstandardbasierter Tests wider.

Einschränkend für die Ergebnisse der Studie II ist zudem in Bezug auf die Stichprobe hinzuzufügen, dass die Schulen zum ersten Messzeitpunkt – im Gegensatz zu den anderen Studien – freiwillig an der VERA-6-Testung teilnehmen konnten. Wir können nicht ausschließen, dass die Freiwilligkeit zur Teilnahme zum Zeitpunkt der 6. Klassenstufe möglicherweise mit Schulmerkmalen zusammenhängt, die sich in den Befunden widerspiegeln. Selektivitätsanalysen mit Hilfe von *funnel plots* geben jedoch hierfür keinen Hinweis, da sie insgesamt zeigen, dass Schulen mit niedriger und hoher Prognosegüte nicht über- bzw. unterrepräsentiert waren (Studie II, ESM 5).

5.3.4 Methode und Ausblick

In allen drei Einzelstudien haben wir grundlegend auf etablierte statistische Verfahren zur Untersuchung unserer Forschungsfragen zurückgegriffen, wie Regressionsanalysen, klassifikatorische Analysen und Meta-Analysen. Allerdings kann das dreischrittige Vorgehen zur Datenauswertung von Cheung und Jak (2016), welches Regressionsanalysen mit Meta-Analysen kombiniert (s. Studie II), als innovativer Ansatz eingestuft werden. Dieses Vorgehen zur Datenauswertung ist vor allem deshalb attraktiv, da es eine flexible Analysemethode zur Untersuchung der Validitätsgeneralisierung verschiedenster Gütekennwerte darstellt – selbst bei sehr umfangreichen Datensätzen. Jedoch muss sich dieses Vorgehen in zukünftigen Studien weiter bewähren, sodass die Grenzen und Umsetzungsvarianten evaluiert werden können. Insbesondere wären zukünftig Studien wünschenswert, die konkrete Hinweise zum Umgang mit fehlenden Werten liefern, da es bisher keine diesbezüglichen Standards bei diesem Vorgehen gibt. Gleiches gilt unter anderem auch für den Umgang mit fehlenden Werten bei klassifikatorischen Analysen (s. Studie III).

Darüber hinaus sollte die bisherige Zusammenfassung unserer Befunde über die drei Einzelstudien hinweg und deren Einordnung in den bisherigen Forschungsstand lediglich als explorativer Zwischenstand angesehen werden. So wäre es wünschenswert, wenn die Ergebnisse unserer Studien zusammen mit dem bisherigen Forschungsstand in einer zukünftigen Metanalyse systematisch zusammengefasst werden. Dadurch wären eine adäquate Abschätzung der mittleren Prognosegüte und deren Generalisierbarkeit über die Einzelstudien hinweg möglich. Zudem könnte der Einfluss potenzieller Moderatorvariablen – in denen sich die Studien unterschieden – auf die Prognosegüte untersucht werden. So variieren unsere drei Studien beispielsweise in Merkmalen, deren Einflussnahme auf die Prognosegüte bereits gut bekannt ist. Dazu gehört (a) die Prognosedauer, (b) der Prognosezeitraum ohne und mit Übergang von der Primar- in die Sekundarstufe, sowie (c) die Berechnung des inkrementellen prognostischen Mehrwerts auf der Grundlage unterschiedlicher zusätzlicher leistungsprädiktiver Merkmale (in Anzahl und Konstrukten). Zudem variieren unserer Studien in zahlreichen anderen Merkmalen, wie beispielsweise der Testdurchführung oder den untersuchten Messinstrumenten, deren Einfluss auf die Prognosegüte im Rahmen einer Meta-Analyse berücksichtigt werden könnte.

Der Wunsch nach einer zukünftigen Meta-Analyse im deutschsprachigen Raum für die systematische Zusammenfassung der bisherigen Befunde zur Prognosegüte standardisierter und bildungsstandardbasierter Leistungstests bringt unweigerlich das Anliegen mit sich, den potenziell bereits vorliegenden oder zukünftig realisierbaren Datenpool in Deutschland besser für diesbezügliche Fragenstellung oder andere Validierungsstudien auszuschöpfen. Bisher werden im Rahmen von VERA bundesweit Leistungsdaten erfasst, die potenziell für forschungsbezogene Fragestellungen aufbereitet und zur Verfügung gestellt werden könnten. So könnte die Güte bildungsstandardbasierter Tests – nicht nur in Bezug auf die Prognosegüte – auf einer soliden und umfassenden Datengrundlage untersucht werden. Durch die zusätzliche Berücksichtigung von Implementationsmerkmalen in Bundesländern oder Schulen könnten so wertvolle Hinweise auf bedeutsame Einflussfaktoren in Bezug auf die Testentwicklung und die Implementation der Tests gewonnen werden, um das diagnostische Potenzial der Tests an den Schulen zu verbessern. Zudem könnten so wissenschaftlich fundierte Kenntnisse gewonnen und den Schulen zur Verfügung gestellt werden, damit diese den zukünftig angedachten Spielraum im Rahmen der angedachten Weiterentwicklung von VERA (KMK, 2012) zielführend umsetzten können.

5.3.5 Validierung und Ausblick

Insbesondere im Rahmen der Studie III, mit der Untersuchung der klassifikatorischen Prognosegüte, haben wir versucht, mögliche Konsequenzen der Testwertinterpretation bildungsstandardbasierter Tests im Sinne des argumentbasierten Ansatzes nach Kane (2013) in die Validierung einfließen zu lassen. So zeigte sich, dass auf der Grundlage bildungsstandardbasierter Tests die Identifikation „gefährdeter“ Kinder verbessert werden kann, und diese somit das Potenzial besitzen, diagnostische Entscheidungen und damit einhergehende Konsequenzen zu verbessern. Wir haben jedoch im Rahmen der Studie III, sowie den anderen beiden Studien nicht den Nachweis erbracht, dass sich auf der Grundlage der untersuchten Testinterpretationen bzw. -nutzung positive Konsequenzen auf die Leistungsentwicklung der Schülerinnen und Schüler mit sich bringt oder negative Konsequenzen reduziert werden. Dieser Punkt ist zumindest im Rahmen des argumentbasierten Ansatzes als eine zentrale Einschränkung unserer bisherigen Testvalidierung zur Prognosegüte zu bewerten. So bleibt offen, ob sich beispielsweise auf Grundlage der Identifizierung die Initiierung von Fördermaßnahmen, oder einer Binnendifferenzierung im Unterricht, förderlich auf die Leistungen der Schülerinnen oder Schüler auswirkt. Aufgrund der bisher vorliegenden empirischen Befunde wäre zu raten, die

Informationen auf der Grundlage bildungsstandardbasierter Tests als relevante Informationen mit diagnostischem Potenzial einzubeziehen – und nicht zu ignorieren. Diesbezügliche Informationen könnten zielgerichtet durch weiterführende Informationen angereichert werden um valide Schlussfolgerungen in Bezug auf instruktionale Entscheidungen, wie beispielsweise der Unterrichtsgestaltung, im Sinne einer formativen Leistungsmessung zu treffen.

Darüber hinaus wird im Zusammenhang mit anderen Rahmenmodellen zur Validierung von Tests zur formativen Leistungsmessung (DiBello, Pellegrino, Gane & Goldman, 2017) empfohlen, auf ein breites Datenspektrum zur Validierung zurückzugreifen. Dazu gehört beispielsweise das Erfassen von prozessrelevanten Daten, welche während der Testung von Schülerinnen und Schülern gesammelt werden (Überblick s. Ercikan & Pellegrino, 2017). Dies können beispielsweise Daten auf der Grundlage von kognitiven Protokollen, log-Dateien oder eye-tracking-Verfahren sein, die während der Testung gewonnen werden. Solche Daten könnten potenziell ebenso im Rahmen der aktuellen Bestrebungen und z. T. bereits realisierten onlinebasierten Umsetzungen von VERA-Tests (KMK, 2012) realisiert werden. Die Informationen könnten nicht nur zur Validierung der Tests genutzt werden, sondern zudem das diagnostische Potenzial bildungsstandardbasierter Tests für die formative Leistungsmessung erhöhen.

5.4 Implikationen für die bildungsstandardbasierte Leistungsmessung an Schulen

5.4.1 Lehrkräfte

Lehrkräfte stehen der Güte bildungsstandardbasierter Tests skeptisch gegenüber (u. a. Ramsteck & Maier, 2015; Wurster et al., 2017). Dies lässt sich z. T. nachvollziehen, wenn man den übersichtlichen Forschungsstand zur Prognosegüte bildungsstandardbasierter Tests rekapituliert. Dieser ist zugleich aber auch nicht weniger umfangreich als der Forschungsstand zu kommerziell erhältlichen standardisierten Schulleistungstests. Die empirische Evidenz unserer drei Einzelstudien untermauert, dass bildungsstandardbasierte Tests, wie beispielsweise VERA-Tests, eine Prognosegüte auf zukünftige schulische Leistungen in Form von späteren Testleistungen und Schulhalbjahresnoten von bis zu 5 Jahren aufweisen können – die vergleichbar zu sein scheint zu bisherigen kommerziell erhältlichen Schulleistungstests. Damit liegt aktuell unseres Wissens nach für bildungsstandardbasierte Tests die solideste empirische Evidenz zur Prognosegüte standardisierter Testverfahren im deutschsprachigen Raum vor. Für kommerziell erhältliche

standardisierte Schulleistungstests basieren die bisherigen Befunde auf deutlich kürzeren Prognosezeiträumen.

Unsere Befunde gehen jedoch noch weiter über die bisherigen Befunde zu standardisierten Leistungstest in Deutschland hinaus. So liefern unsere Studien empirische Hinweise für einen substantiellen inkrementellen prognostischen Mehrwert bildungsstandardbasierter Tests, neben der Berücksichtigung von Schul(halbjahres)noten über Prognosen von bis zu 5 Jahren. Dies ist durchaus bemerkenswert, wenn man berücksichtigt, dass es sich bei der Testleistung um eine einmalige Leistungsmessung im Vergleich zu den Schul(halbjahres)noten handelt. Die Halbjahresnoten setzen sich aus zahlreichen Leistungsmessungen zusammen und sollten dadurch reliabler sein. Diese Befunde verdeutlichen, dass die Tests neben etablierten Leistungsinformationen an Schulen durchaus relevante Leistungsinformationen liefern können, die die Leistungsdiagnostik an den Schulen verbessern können.

Darüber hinaus zeigte sich aber auch, dass das diagnostische Potenzial bildungsstandardbasierter Tests in Bezug auf die Prognosegüte z. T. substantiell zwischen Schulen (insbesondere für Schulen mit mehreren Bildungsgängen in Hinblick auf den inkrementellen prognostischen Mehrwert) variieren können. Folglich könnten aktuelle Bestrebungen der KMK zur Weiterentwicklung von VERA (KMK, 2012) durchaus zu einer Verbesserung des diagnostischen Potenzials an Schulen führen (bspw. durch Anpassung der Testhefte an das Leistungsniveau der Schülerinnen und Schüler) und somit Lehrkräfte diesen Spielraum zur Verbesserung der Diagnostik an ihren Schulen zielführend nutzen sollten.

Abschließend zeigte sich, dass bildungsstandardbasierte Tests im Sinne eines Screenings durchaus Potenzial zur Identifizierung von Förderbedarf aufweisen und vor allem die Kombination mit Noten die Diagnostik weiter verbessern können. Jedoch sollten in Abhängigkeit der diagnostischen Zielstellung unterschiedliche Strategien zur Kombination der leistungsrelevanten Informationen (s. Studie III; UND-Strategie, ODER-Strategie) und möglicherweise ein anderer Schwellenwert (bspw. Orientierung am Mindest- oder Regelstandard) gewählt werden, da sich diese auf die Klassifikationsgüte auswirken. Eine Möglichkeit zur Abschätzung des Nutzens verschiedener Screeningverfahren auf der Basis kontextueller Faktoren an Schulen (wie der Basisrate) liefert zum Beispiel VanDerHeyden (2013). Aus unseren diesbezüglichen Befunden zur klassifikatorischen Prognosegüte wurde zudem veranschaulicht, dass die Berücksichtigung mehrerer Informationsquellen die Leistungsdiagnostik verbessern kann. Die Übereinstimmung von diagnostischen

Informationen verschiedener Informationsquellen stellt dabei einen zusätzlichen diagnostischen Mehrwert für die Diagnostik dar. Folglich sollten Übereinstimmungen nicht dazu führen, dass diese unberücksichtigt bleiben, wie sich dies z. T. als Grund von Lehrkräften für die Nicht-Nutzung von VERA-Daten andeutet (Ramsteck & Maier, 2015). Lehrkräfte sollten unabhängig hiervon bildungsstandardbasierte Testinformationen als relevante Informationen neben den zahlreichen weiteren leistungsbezogenen Informationen im Schulkotext ansehen und berücksichtigen. Dabei sollte das Potenzial der Tests für die Individualdiagnostik jedoch nicht überschätzt werden. Somit empfehlen wir nicht die Testleistungen als alleinige Informationsquelle zur Ableitung von Förderbedarf von Schülerinnen und Schüler zu nutzen. Vielmehr könnten auf der Grundlage der Testergebnisse in Kombination mit Schulnoten (oder anderen leistungsrelevanten Informationen) diagnostische Entscheidungen zum Förderbedarf abgeleitet werden. Entsprechende Entscheidungen sollten im Sinne einer formativen Leistungsdiagnostik (bspw. durch weitere Tests und Beobachtungen im Unterricht) fortlaufend überprüft werden.

5.4.2 Bildungsadministration

Aktuelle Bestrebungen im Rahmen von VERA (KMK, 2012). zur Modularisierung und Flexibilisierung des Einsatzes von VERA-Tests an Schulen fördert die Autonomie der Schulen und unterstützt damit die Nutzung von VERA zur formativen Leistungsmessung. Schulen haben so die Möglichkeit, die Sammlung von Leistungsdaten auf ihre fokussierte Zielstellung auszurichten. Unsere Befunde (aus Studie II) liefern empirische Hinweise, dass sich dadurch das diagnostische Potenzial der Tests an den einzelnen Schulen verbessern lässt. Jedoch darf in diesem Zusammenhang nicht vergessen werden, dass Schulen auch valide Informationen benötigen, um diesen Handlungsspielraum zielführend nutzen zu können. Hierfür sollte anwendungsorientierte Forschung im Rahmen bildungsstandardbasierter Testungen, u. a. im Rahmen von VERA, zielgerichtet unterstützt und ausgebaut werden. Ein wichtiger Anfangsschritt wäre meines Erachtens, die bisher jährlich generierten Daten systematisch zu Längsschnittdatensätzen zusammenzufassen (wofür es bereits praktikable Beispiele in Bundesländern unter Wahrung des Datenschutzes gibt) und unabhängigen Forscherinnen und Forschern über Forschungsdatenzentren zur Verfügung zu stellen. So könnten Forschungsfragen zur Güte der Tests untersucht werden, die keine weiteren separaten Einzelerhebungen an Schulen notwendig machen würden. Möglicherweise ließen sich diese Datensätze mit geringfügigen Informationen der Schulen zur Implementation von VERA

anreichern und somit das Potenzial zur Untersuchung weiterer Forschungsfragen ausbauen. Der diesbezügliche Aufwand könnte sowohl für Schulen als auch für Forscherinnen und Forscher gering gehalten werden, da sich diese an die regulären Erhebungen anknüpfen ließen – die auch unabhängig von der Nutzung im Rahmen anwendungsorientierter Forschung – jährlich an den Schulen stattfinden und somit keine zusätzlichen Erhebungen für Schulen darstellen. Dadurch ließen sich bereits bestehende Datenbestände zur Generierung weiterer Erkenntnisse zur Verbesserung der Güte bildungsstandardbasierter Tests und deren Implementation ausschöpfen.

Außerhalb des Kontextes von VERA wäre im Sinne einer formativen Leistungsmessung zu überlegen, ob Lehrkräfte an Schulen außerhalb der VERA-Testungen besser unterstützt werden sollten, eine eigenständige Überprüfung des Lernstandes ihrer Schülerinnen und Schüler vornehmen zu können. Dies könnte beispielsweise durch die Bereitstellung computerbasierter bildungsstandardbasierter Testhefte erfolgen, die zudem eine adäquate Auswertung und Rückmeldung realisieren. Auf diese Weise hätten Lehrkräfte die Möglichkeit im Sinne einer Lernverlaufsdiagnostik, die Entwicklung der Leistungen ihrer Schülerinnen und Schüler in Bezug auf die Bildungsstandards zu überprüfen und somit Rückmeldungen zur Auswirkung instruktionaler Anpassungen der Schul-, Klassen-, oder Individualebene auf die Leistungsentwicklung ihrer Schülerinnen und Schüler in Bezug auf die Bildungsstandardmetrik zu erhalten. Im Rahmen der VERA-Testungen erhalten Lehrkräfte lediglich einmalig zu den gleichen Schülerinnen und Schülern eine Einordnung des Leistungsstandes auf der Bildungsstandardmetrik, die eine grobe Einordnung darstellt und somit zur Lernverlaufsdiagnostik – im Sinne formativer Leistungsmessung – nur sehr eingeschränkt nutzbar ist. Auf diese Weise könnte beispielsweise zumindest einmal im Jahr eine Überprüfung der Leistungen in Hinblick auf die Bildungsstandardmetrik realisiert werden. Neben VERA-Tests liegen auch kommerzielle bildungsstandardbasierte Tests vor, die Lehrkräfte potenziell zur Realisierung einer diesbezüglichen Lernverlaufsdiagnostik nutzen könnten. Jedoch lässt sich auf der Grundlage einer Internetrecherche schlussfolgern, dass diese Bereitstellung zu den Vorjahren abgenommen hat und aktuell nur noch sehr eingeschränkt bildungsstandardbasierte Testhefte kommerziell erworben werden können. Zwar können aktuell Lehrkräfte auch über (kostenfreie) Aufgabendatenbanken ((Institut für Schulqualität der Länder Berlin und Brandenburg e.V. [ISQ], n. d.) Testhefte mit bildungsstandardbasierten Aufgaben zusammenstellen, aber dennoch kann so keine

raschkonforme Auswertung der Daten realisiert werden, die für eine adäquate Leistungsbestimmung der Schülerleistungen notwendig wäre.

5.4.3 Forschung

Der Überblick zum Forschungsstand zur Prognosegüte bildungsstandardbasierter und speziell kommerziell erhältlicher standardisierter Leistungstests in Deutschland hat deutlich gemacht, dass in bisherigen Studien der Nachweis eines inkrementellen prognostischen Mehrwerts von Leistungstests weitestgehend unberücksichtigt bleibt. Diese empirische Evidenz ist jedoch wichtig, um die praktische Relevanz der Leistungstests neben bereits bestehenden Verfahren abschätzen zu können. Dadurch wird der Praxisbezug stärker im Validierungsprozess der Leistungstests berücksichtigt und die Generalisierbarkeit der Befunde auf die Schulpraxis erhöht. Zudem konnte dabei nicht nur verstärkt berücksichtigt werden, wie valide ein bestimmtes Verfahren ist, sondern auch wie valide ein Verfahren in Kombination mit anderen Verfahren zur Leistungsmessung ist. Lehrkräfte werden aktuell aufgrund des national und international verfolgten datenbasierten Ansatzes zur Verbesserung der Leistungen ihrer Schülerinnen und Schüler (Ikemoto & Marsh, 2007; Mandinach, 2012) mit zahlreichen leistungsrelevanten Informationen auf der Grundlage verschiedenster Verfahren konfrontiert. So stellt sich nicht unbedingt die Frage, welches Testverfahren genutzt werden sollte, sondern vielmehr, wie die Informationen verschiedener Verfahren zielführend kombiniert werden sollten um eine valide Entscheidung zu treffen.

Zudem wäre eine zukünftige Meta-Analyse wünschenswert, die eine adäquate Zusammenfassung der bisherigen Befunde zur Prognosegüte standardisierter und bildungsstandardbasierter Tests ermöglichen würde. Im Rahmen dieser Meta-Analyse könnte beispielsweise der Einfluss bestimmter Durchführungsmodalitäten auf die Güte der Tests untersucht werden und deren Relevanz in Bezug auf andere Faktoren eingeordnet werden.

Abschließend soll hier auf das Potenzial des meta-analytischen Ansatzes nach Cheung und Jak (2016 s. Studie II) zur Validitätsgeneralisierung hingewiesen werden. Mit diesem Ansatz lässt sich die Heterogenität zwischen verschiedensten Kontexten (bspw. Schulen, Klassen, Bezirken) in Bezug auf verschiedenste Gütekennwerte von Leistungstests (bspw. Reliabilität) – auch für sehr große Datensätze – flexibel untersuchen. Der Erkenntnisgewinn aus diesbezüglicher Forschung könnte im Kontext von VERA-Testungen zu einer besseren

Zielformulierung des Einsatzes solcher Tests beitragen und empirische Hinweise für bedeutsame Ansatzpunkte zur Förderung der Diagnostik an Schulen liefern.

Letztlich untermauert die starke Kohärenz der Befunde zur Prognosegüte über unsere drei Studien hinweg, dass auf der Grundlage bildungsstandardbasierter Testergebnisse späterer Schulerfolg prognostiziert werden kann. Somit scheinen die Bildungsstandards auf der Grundlage des bisherigen Forschungsstandes den Anspruch zu erfüllen, Leistungen zu definieren, die für anschlussfähiges Lernen von Bedeutung sind und infolgedessen erfüllen sie wichtige Schlüsselmerkmale einer formativen Leistungsmessung.

5.5 Fazit

Auf der Grundlage dreier Längsschnittstudien ergab sich auf einer sehr umfassenden Datenbasis ein kohärentes Bild zur Güte bildungsstandardbasierter Tests zur Prognose späterer bildungsstandardbasierter Testleistungen und Schulnoten von bis zu 5 Jahren. Dieses legt nahe, dass bildungsstandardbasierte Tests, wie diese beispielsweise im Rahmen von VERA-Tests jährlich bundesweit an allen öffentlichen Schulen in der 3. und 8. Jahrgangsstufe eingesetzt werden, eine vergleichbare Prognosegüte zu kommerziell erhältlichen standardisierten Schulleistungstests aufweisen. Zudem zeigte sich, dass bildungsstandardbasierte Tests einen prognostischen Mehrwert neben weiteren etablierten leistungsrelevanten Merkmalen, wie der Schulhalbjahresnoten, besitzen können. Die Ausprägung der Prognosegüte bildungsstandardbasierter Tests variiert allerdings zum Teil bedeutsam zwischen Einzelschulen. Diesbezügliche Befunde untermauern die notwendige Berücksichtigung schulspezifischer Merkmale, um eine hinreichende Güte der Tests an allen Schulen gewährleisten zu können. Darüber hinaus liegen nun auch erste empirische Hinweise vor, die eine hinreichende Güte von VERA-Tests auf Individualebene zur Prognose zukünftiger Schülerleistungen im Sinne eines Screenings stützen. Auf der Grundlage dieses Forschungsstandes zur Prognosegüte scheinen die Bildungsstandards dem an sie gestellten Anspruch gerecht zu werden: Leistungen zu definieren, die für anschlussfähiges Lernen von Bedeutung sind. Zukünftig sollte geprüft werden, inwiefern das diesbezügliche Potenzial bildungsstandardbasierter Tests zur Prognose des Schulerfolgs zur Förderung der Leistungsentwicklung von Schülerinnen und Schüler an Schulen beiträgt und in Zusammenhang mit anderen relevanten Informationen zielführend ergänzt werden kann.

6 Literatur

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Amrein-Beardsley, A., Berliner, D. C. & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, 18* (14), 1–36. https://doi.org/10.14507/epaa.v18n14.2010

Baumert, J., Lüdtke, O., Trautwein, U. & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, *4* (3), 165–176. https://doi.org/10.1016/j.edurev.2009.04.002

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21* (1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G. & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100* (2), 431–449. https://doi.org/10.1037/a0038047

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. et al. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, *277* (3), 826–832. https://doi.org/10.1148/radiol.2015151516

Cheung, M. W.-L. & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 1–19. https://doi.org/10.3389/fpsyg.2016.00738

Cicmanec, K. M., Johanson, G. & Howley, A. (2001). *High school mathematics teachers: Grading practice and pupil control ideology.* Verfügbar unter: http://files.eric.ed.gov/fulltext/ED453290.pdf

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3. Auflage). Mahwah, NJ: Lawrence Erlbaum Associates.

Cumming, G. & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. New York: Routledge, Taylor & Francis Group.

Dahlke, J. A., Kostal, J. W., Sackett, P. R. & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. Advance online publication. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000316

Decker, D. M. & Bolt, S. E. (2008). Challenges and opportunities for promoting student achievement through large-scale assessment results: Research, reflections, and future directions. *Assessment for Effective Intervention, 34* (1), 43–51. https://doi.org/10.1177/1534508408314173

DiBello, L. V., Pellegrino, J. W., Gane, B. D. & Goldman, S. R. (2017). The contribution of student response processes to validity analyses for instructionally supportive assessments (NCME applications of educational measurement and assessment book series). In K. Ercikan & J.W. Pellegrino (Hrsg.), *Validation of score meaning for the next generation of assessments: The use of response processes* (S. 85–99). New York, N.Y: Routledge.

Ercikan, K. & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels. Guidelines for assessment design. *Applied Measurement in Education, 15* (3), 269–294. https://doi.org/10.1207/S15324818AME1503_3

Ercikan, K. & Pellegrino, J. W. (Hrsg.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes* (NCME applications of educational measurement and assessment book series). New York, N.Y: Routledge.

Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Los Angeles, CA: SAGE.

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K. & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78* (4), 423–445.

Götz, L., Lingel, K., Schneider, W., Götz, L., Lingel, K. & Schneider, W. (2013a). *DEMAT 6+: Deutscher Mathematiktest für sechste Klassen*. Göttingen: Hogrefe.

Götz, L., Lingel, K., Schneider, W., Götz, L., Lingel, K. & Schneider, W. (2013). *DEMAT 5+: Deutscher Mathematiktest für fünfte Klassen*. Göttingen: Hogrefe.

Graf, T., Emmrich, R., Harych, P. & Brunner, M. (2013). Durchführungseffekte bei Vergleichsarbeiten in Jahrgangsstufe 8. *Empirische Pädagogik*, *27* (4), 459–473.

Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie*, *30* (4), 201–211. https://doi.org/10.1024/1010-0652/a000182

Groß Ophoff, J. (2013). Der Effekt der Bezugsnormorientierung auf die Reflexion und Nutzung von Rückmeldungen aus Vergleichsarbeiten. *Empirische Pädagogik, 27* (4), 442–458.

Hellrung, K. & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F.E. Weinert (Hrsg.), *Enzyklopädie der Psychologie: Psychologie des Unterrichts und der Schule* (Band 3, S. 71–176). Göttingen: Hogrefe.

Hildebrandt, J. & Watermann, R. (2015). Prognostische Validität von curricular gemessenen Testleistungen am Ende der Grundschulzeit. Vortrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF) in Bochum (11.-13. März) im Rahmen des Symposiums „Heterogenität standardisierter Testverfahren im Kontext der KMK-Gesamtstrategie zum Bildungsmonitoring: Validität. Wert. Schätzen.". Bochum.

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe*. Münster: Waxmann.

Ikemoto, G. & Marsh, J. A. (2007). Cutting through the „data-driven“ mantra: Different conceptions of data-driven decision making. In P.A. Moss (Hrsg.), *Evidence and decision making* (S. 104–131). Blackwell Publishing.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Aufl.). Weinheim: Beltz Verlag.

Institut für Schulqualität der Länder Berlin und Brandenburg e.V. (ISQ). (n. d.). *Aufgabenbrowser*. Verfügbar unter: https://www.aufgabenbrowser.de/itemdb/login.seam

Institut zur Qualitätsentwicklung im Bildungswesen. (n. d.). VERA-3 Testdomänen 2006 bis 2020 und VERA-8 Testdomänen 2009 bis 2020. Verfügbar unter: https://www.iqb.hu-berlin.de/vera/aktuell

Isaac, K. (2013). Kriteriale Bezugsnormen bei Lernstanderhebungen. Bilanz und Perspektiven. *Empirische Pädagogik, 27* (4), 407–422.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112* (3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R.W. Lissitz (Hrsg.), *The concept of validity: revisions, new directions, and applications* (S. 39–64). Charlotte, NC: Information Age Pub.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50* (1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2017). Using empirical results to validate preformance standards. In S. Blömeke & J.-E. Gustafsson (Hrsg.), *Standard setting in education* (S. 11–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50856-6_2

Klauer, K. J. (1987). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.

Köller, O. (2005). Formative assessment in classrooms: A review of the empirical german literature. In Organization of Economic Co-operation and Development (Hrsg.), *Formative assessment: Improving learning in secondary classrooms* (S. 265–279). Paris: OECD Publishing.

Köller, O. & Reiss, K. (2013). Mathematische Kompetenzen messen: Gibt es Unterschiede zwischen standardisierten Verfahren und diagnostischen Tests? (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, A. Heinze, W. Schneider & U. Trautwein (Hrsg.), *Diagnostik mathematischer Kompetenzen* (Band 11, S. 25–37). Göttingen: Hogrefe.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22* (2), 18–26. https://doi.org/10.1111/j.1745-3992.2003.tb00124.x

Krajewski, K., Liehm, S. & Schneider, W. (2004). *DEMAT 2+. Deutscher Mathematiktest für zweite Klassen*. Göttingen: Beltz.

Kulow, A.-C. (2011). Rechtliche Spielräume und Grenzen der Leistungsüberprüfung und Leistungsbewertung (Professionswissen für Lehrerinnen und Lehrer). In W. Sacher & F. Winter (Hrsg.), *Diagnose und Beurteilung von Schülerleistungen: Grundlagen und Reformansätze* (Band 4, S. 73–106). Baltmannsweiler: Schneider-Verl. Hohengehren.

Kultusministerkonferenz. (n. d.). *Bildungsstandards*. Verfügbar unter: https://www.kmk.org/dokumentation-statistik/beschluesse-und-veroeffentlichungen/bildung-schule/qualitaetssicherung-in-schulen.html#c2365

Kultusministerkonferenz. (2004). *Bildungsstandards der Kultusministerkonferenz: Erläuterungen zur Konzeption und Entwicklung (Am 16.12.2004 von der Kultusministerkonferenz zustimmend zur Kenntnis genommen)*. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Konzeption-Entwicklung.pdf

Kultusministerkonferenz. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Neuwied: Lichterhand. Verfügbar unter: https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/Beschluesse_Veroeffentlichungen/Bildungsmonitoring_Broschuere_Endf.pdf

Kultusministerkonferenz. (2012). *Vereinbarung zur Weiterentwicklung von VERA (Beschluss der Kultusministerkonferenz vom 08.03.2012 i.d. F. vom 10.03.2018).* Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Kultusministerkonferenz. (2013a). *Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich „Lesen – mit Texten und Medien umgehen“ – Primarbereich – (Auf Grundlage des Ländervergleichs 2011 überarbeiteter Entwurf*

in der Version vom 13. Februar 2013). Verfügbar unter: https://www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Deutsch_L_2.pdf

Kultusministerkonferenz. (2013b). *Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Mathematik. Beschluss der Kultusministerkonferenz vom 20./21.10.2011. (Auf Grundlage des Ländervergleichs 2012 überarbeitete Version in der Fassung vom 11.10.2013).* Verfügbar unter: https://www.iqb.hu-berlin.de/bista/ksm

Kultusministerkonferenz. (2013c). *Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Primarbereich (Jahrgangsstufe 4) (Auf Grundlage des Ländervergleichs 2011 überarbeitete Version in der Fassung vom 11. Februar 2013).* Verfügbar unter: http://www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Mathemati_2.pdf

Kultusministerkonferenz. (2014). *Integriertes Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Deutsch für den Kompetenzbereich Lesen – mit Texten und Medien umgehen. Beschluss der Kultusministerkonferenz (KMK) vom 11.12.2014.* Verfügbar unter: https://www.iqb.hu-berlin.de/bista/ksm

Kultusministerkonferenz (Hrsg.). (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Berlin: Kluwer. Verfügbar unter: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

Kultusministerkonferenz. (2016). *KMK Bildungsmonitoring (II) Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Carl Link. Verfügbar unter: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

Kuper, H., Maier, U., Graf, T., Muslic, B. & Ramsteck, C. (2017). Datenbasierte Schulentwicklung mit Vergleichsarbeiten aus der Perspektive von Lehrkräften, Fachkonferenzleitungen, Schulleitungen und Schulaufsichten – Qualitative Fallstudien aus vier Bundesländern. In Bundesministerium für Bildung und Forschung. (Hrsg.), *Steuerung im Bildungssystem Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (Band 43, S. 39–67). Berlin.

Lenhard, W. & Schneider, W. (2006). *Ein Leseverständnistest für Erst- bis Sechstklässler (ELFE 1–6)*. (M. Hasselhorn, H. Marx & W. Schneider, Hrsg.). Göttingen: Hogrefe.

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2008). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (S. 149–167). Wiesbaden: VS Verlag für Sozialwissenschaften. Verfügbar unter: http://link.springer.com/chapter/10.1007/978-3-531-90865-6_9

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: Verlag für Sozialwissenschaften.

Maier, U. (2010). Vergleichsarbeiten im Spannungsfeld zwischen formativer und summativer Leistungsmessung. *DDS - Die Deutsche Schule, 102* (1), 60–69.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47* (2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Marx, H. (1992). Frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten: Bestandsaufnahme bisheriger Bewertungsgesichtspunkte von Längsschnittstudien. *Zeitschrift für Pädagogische Psychologie*, *6* (1), 35–48.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50* (9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nachtigall, C. (2014). *Landesbericht. Thüringer Kompetenztests 2014*. Verfügbar unter: https://www.kompetenztest.de/downloads/kompetenztests

Nagy, G., Haag, N., Lüdtke, O. & Köller, O. (2017). Längsschnittskalierung der Tests zur Überprüfung des Erreichens der Bildungsstandards der Sekundarstufe I im PISA-Längsschnitt 2012/2013. *Zeitschrift für Erziehungswissenschaft, 20* (2), 259–286. https://doi.org/10.1007/s11618-017-0755-1

OECD. (2016). *PISA 2015 results (volume II): Policies and practices for successful schools* (PISA). (Y. Belfali, Programme for International Student Assessment & Organisation for Economic Co-operation and Development, Hrsg.). Paris: OECD.

Pant, H. A., Emmrich, R., Harych, P. & Kuhl, P. (2011). Leistungsüberprüfung durch Schulleistungsstudien und Vergleichsarbeiten (Professionswissen für Lehrerinnen und Lehrer). In W. Sacher & F. Winter (Hrsg.), *Diagnose und Beurteilung von Schülerleistungen: Grundlagen und Reformansätze* (Band 4, S. 123–141). Baltmannsweiler: Schneider-Verl. Hohengehren [u.a.].

Pant, H. A., Tiffin-Richards, S. P. & Stanat, P. (2017). Standard setting: Bridging the worlds of policy making and research. In S. Blömeke & J.-E. Gustafsson (Hrsg.), *Standard Setting in Education* (S. 49–68). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50856-6_4

Pellegrino, J. W., DiBello, L. V. & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51* (1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Petscher, Y., Kim, Y.-S. & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention, 36* (3), 158–166. https://doi.org/10.1177/1534508410396698

Ramsteck, C. & Maier, U. (2015). Testdatenbasierte Schul- und Unterrichtsentwicklung. Analyse von Handlungsmustern bei der Rezeption und Nutzung von Vergleichsarbeitsdaten. In J. Schrader, J. Schmid, K. Amos & A. Thiel (Hrsg.), *Governance von Bildung im Wandel* (S. 119–144). Wiesbaden: Springer Fachmedien Wiesbaden. Verfügbar unter: http://link.springer.com/10.1007/978-3-658-07270-4_6

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L. & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47* (6), 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

Richter, D. & Böhme, K. (2014). Vergleichsarbeiten im Fokus: Welche Funktionen erfüllt der Test aus Sicht von Lehrkräften? *Schulmanagement, 45* (2), 12–14.

Roick, T., Gölitz, D. & Hasselhorn, M. (2004). *DEMAT 3+. Deutscher Mathematiktest für dritte Klassen*. Göttingen: Beltz.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R. & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135* (1), 1–22. https://doi.org/10.1037/a0013978

Schmid, C., Paasch, D. & Katstaller, M. (2016). Kompositionseffekte bei der Notenvergabe in Mathematik auf der 4. Schulstufe der österreichischen Volksschule. *Zeitschrift für Bildungsforschung, 6* (3), 265–283. https://doi.org/10.1007/s35834-016-0170-3

Schmidt, S., Ennemoser, M., Krajewski, K., Schmidt, S., Ennemoser, M. & Krajewski, K. (2013). *DEMAT 9: Deutscher Mathematiktest für neunte Klassen mit Ergänzungstest Konventions- und Regelwissen*. Göttingen: Hogrefe.

Schreiner, C., Breit, S. & Haider, G. (2008). Zur Validität der Mathematiknoten. Ein Vergleich von Lehrerbeurteilung und Leistungsmessung bei PISA. In F. Hofmann, C. Schreiner & J. Thonhauser (Hrsg.), *Qualitative und quantitative Aspekte: zu ihrer Komplementarität in der erziehungswissenschaftlichen Forschung* (S. 211–223). Münster: Waxmann.

Schütze, B., Souvignier, E. & Hasselhorn, M. (2018). Stichwort – Formatives assessment. *Zeitschrift für Erziehungswissenschaft, 21* (4), 697–715. https://doi.org/10.1007/s11618-018-0838-7

Sedlmeier, P. & Renkewitz, F. (2013). *Forschungsmethoden und Statistik. Ein Lehrbuch für Psychologen und Sozialwissenschaftler.* (2., aktualisierte und erw. Aufl.). München [u.a.]: Pearson.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont: Cengage Learning.

Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik, 35* (1), 79–99. https://doi.org/10.1007/s13138-013-0056-z

Statistisches Bundesamt (Destatis). (2017). *Bildung und Kultur: Allgemeinbildende Schulen (Schuljahr 2016/2017).* Fachserie 11 No. 2110100177004. Verfügbar unter: https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/AllgemeinbildendeSchulen2110100177004.pdf?__blob=publicationFile

Tiffin-Richards, S. P. (2011). *Setting standards for the assessment of english as a foreign language. Establishing validity evidence for criterion-referenced interpretations of test-scores*. Berlin: Freie Universität.

Toulmin, S. (1958). *The uses of argument*. Cambridge England: Cambridge University Press.

Toulmin, S. E. (2001). *Return to reason*. Cambridge, Mass.: Harvard Univ. Press.

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42* (4), 402–414.

Winkelmann, H., Robitzsch, A., Stanat, P. & Köller, O. (2012). Mathematische Kompetenzen in der Grundschule. Struktur, Validierung und Zusammenspiel mit allgemeinen kognitiven Fähigkeiten. *Diagnostica*, *58* (1), 15–30. https://doi.org/10.1026/0012-1924/a000061

Wurster, S., Bach, A., Schliesing, A., Thillmann, K., Pant, H. A. & Thiel, F. (2017). Schulen als Steuerungsakteure im Bildungssystem – datenbasierte Schul- und Unterrichtsentwicklung aus der Perspektive von Schulleitungen, Fachkonferenzleitungen und Lehrkräften. In Bundesministerium für Bildung und Forschung (BMPF) (Hrsg.), *Steuerung im Bildungssystem Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (Band 43, S. 177–207). Berlin.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31* (6), 412–422. https://doi.org/10.1177/0741932508327463

Anhang A: Manuskript und elektronische Supplemente - Studie I

STUDIE I: Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen?

Die Teilstudie ist als Zeitschriftenbeitrag veröffentlicht und wie folgt zugänglich:

Fuchs, G. & Brunner, M. (2017). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie, 31* (1), 27-39. https://doi.org/10.1024/1010-0652/a000195

Zusammenfassung

Bildungsstandardbasierte Tests werden bundesweit in der Primar- und Sekundarstufe zur Überprüfung von fachlichen Kompetenzen der Schülerinnen und Schüler in Deutschland eingesetzt, um ausgehend vom aktuellen Leistungsstand auch über deren zukünftigen schulischen Erfolg zu informieren. Wie gut diese Prognose gelingt, untersucht die vorliegende Längsschnittstudie anhand bildungsstandardbasierter Tests zur Erfassung der mathematischen Kompetenz und der Lesekompetenz in Deutsch für Brandenburger Grundschulkinder in der 3. bzw. 4. Klasse. Die Ergebnisse aus Regressionsanalysen zeigen, dass bildungsstandardbasierte Kompetenztests Prognosen auf 2 bzw. 3 Jahre später erhobene Kompetenztestergebnisse ($\beta \geq .56$) sowie Schulnoten ($\beta \leq -.47$) in vergleichbarem Ausmaß wie kommerziell erhältliche, standardisierte Schulleistungstests ermöglichen. Zudem ließ sich für beide Fächer eine substanzielle Vorhersagekraft der Kompetenztests für die Gymnasialempfehlung am Ende der 6. Klasse nachweisen. Auch bei Kontrolle in den Vorhersagemodellen für weitere leistungsrelevante Merkmale (Schulnoten, Intelligenz und familiärer Hintergrund), blieb ein prognostischer Mehrwert der Kompetenztests für alle untersuchten Kriterien des schulischen Erfolgs bestehen.

Schlüsselwörter: Bildungsstandards; Grundschule; (prognostische) Validität; Mathematik; Lesekompetenz (Deutsch)

Abstract

Standard-based tests are used in Germany to assess primary and secondary students' domain-specific competencies where students' current proficiency levels are also supposed to inform about their future school success. The present longitudinal study examines the extend to which standard-based test scores that measure proficiency in mathematics and German reading comprehension (in 3rd and 4th grade) predict school success for children at primary schools in the federal state of Brandenburg. Regression analyses show that standard-based test scores predicted test results ($\beta \geq .56$) and grades ($\beta \leq -.47$) two to three years later. The predictive power was comparable to that of commercial achievement tests. Further, standard-based tests in both domains predicted teachers' tracking recommendations for attending the highest academic track at the end of 6th grade. The predictive power for all criteria of school success remained even when controlling for additional student characteristics (e.g., grades, intelligence, socio-economic background).

Keywords: standard of education; primary school; (prognostic) validity; mathematics achievement; german reading comprehension

Bildungsstandards beschreiben „die fachbezogenen Kompetenzen einschließlich zugrunde liegender Wissensbestände, die Schülerinnen und Schüler bis zu einem bestimmten Zeitpunkt ihres Bildungsganges erreicht haben sollen“ (Kultusministerkonferenz [KMK], 2006, S. 6). Zur Überprüfung der Bildungsstandards werden bundesweit Kompetenztests im Rahmen des IQB-Ländervergleichs und der Vergleichsarbeiten (VERA) eingesetzt (KMK, 2015). So gibt es beispielsweise VERA 3 seit dem Schuljahr 2008/2009, und seit 2012 sind bundesweit alle Lehrkräfte der dritten Jahrgangsstufe an öffentlichen Schulen verpflichtet, jährlich in mindestens einem Fach VERA 3 durchzuführen (KMK, 2012). Mit dem Einsatz bildungsstandardbasierter Tests wird das übergeordnete Ziel einer datengestützten Qualitätsentwicklung und -sicherung im Bildungswesen verfolgt (KMK, 2006). Essenziell hierfür ist eine hohe psychometrische Qualität der Tests bzw. der gewonnenen Daten, da weitreichende Entscheidungen auf Grundlage dieser Informationen getroffen werden. Mit dem Nachweis einer hohen Datenqualität wird zum einen der Forderung einschlägiger Teststandards (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999) nachgekommen. Da VERA-Tests in einigen pädagogischen Kreisen heftig kritisiert werden (u. a. Kuhn, 2014), ist damit zum anderen auch die Hoffnung verbunden, dass Lehrkräfte sich intensiver mit den Ergebnisrückmeldungen aus VERA auseinandersetzen. Die Ergebnisrückmeldungen bei VERA informieren Lehrkräfte über den aktuellen Leistungsstand ihrer Schülerinnen und Schüler und geben Hinweise für deren weitere schulische Entwicklung. Denn die Bildungsstandards „formulieren fachliche und fachübergreifende Basisqualifikationen, die für die weitere schulische und berufliche Ausbildung von Bedeutung sind und die anschlussfähiges Lernen ermöglichen“ (KMK, 2004, S. 7). Dieser Anspruch impliziert, dass auf Grundlage der Ergebnisse bei bildungsstandardbasierten Tests zentrale Kriterien des Schulerfolgs vorhersagbar sein sollten. Trotz ihrer großen Bedeutung für die Bildungspolitik und Schulpraxis existieren bislang jedoch kaum Studien zur Vorhersagegüte bildungsstandardbasierter Tests für den Grundschulbereich. Ziel der vorliegenden Arbeit ist deshalb die Analyse der Prognosekraft bildungsstandardbasierter Tests in Mathematik und Deutsch (Lesekompetenz) für zentrale Schulerfolgskriterien (Testleistungen, Schulnoten, Gymnasialempfehlung) für Kinder an Brandenburger Grundschulen von der dritten bzw. vierten bis zur sechsten Klasse.

Forschungsstand zur Prognosekraft von schulischen Leistungstests

In diesem Abschnitt fokussieren wir auf Zusammenhänge bildungsstandardbasierter Tests der Mathematikkompetenz und der Lesekompetenz (in Deutsch) mit drei wichtigen Kriterien des Schulerfolgs: (a) Testleistungen und (b) Schulnoten desselben Faches und (c) Gymnasialempfehlung. Von besonderer Bedeutung ist dabei auch, inwiefern bildungsstandardbasierte Tests einen prognostischen Mehrwert (inkrementelle Validität) gegenüber anderen leistungsrelevanten Prädiktoren (z. B. Schulnoten, Intelligenz, sozioökonomischer Status) besitzen. Ergebnisse dieser Analysen sind hochrelevant für die Akzeptanz bildungsstandardbasierter Tests in der Praxis. Für Lehrkräfte, denen in der Regel lediglich Schulnoten vorliegen, kann der Nachweis eines diagnostischen Mehrwerts von bildungsstandardbasierten Tests zu einem wichtigen Argument werden, um den Aufwand für die Durchführung dieser Tests zu rechtfertigen. Da für einige Fragen zur Prognosekraft bildungsstandardbasierter Tests noch keine relevanten Ergebnisse vorliegen, fassen wir zur Einordnung der vorliegenden Befunde auch die Befundlage für kommerzielle, standardisierte Schulleistungstests zusammen, die frei für Lehrkräfte erhältlich sind. (Eine detaillierte Auflistung dieser Ergebnisse befindet sich im Elektronischen Supplement [ES] ESM-1.)

Prognose von Testleistungen

Längsschnittliche Befunde für die Vorhersagekraft von bildungsstandardbasierten Testleistungen auf zukünftige Testleistungen berichtet Nachtigall (2014) für Mathematik und für Deutsch. Die durchschnittlichen Korrelationen liegen zwischen $r = .63$ und $r = .67$ für ein Vorhersageintervall von der dritten zur sechsten Jahrgangsstufe bzw. von der dritten zur achten Jahrgangsstufe. Diese Ergebnisse liegen in einem ähnlichen Bereich wie auch andere standardisierte Leistungstests. Hier liegen die Korrelationen zwischen $.55 \leq r \leq .80$ in Mathematik und $.31 \leq r \leq .75$ für Deutsch-Lesen. Gründe für die Variabilität der Korrelationen sind u. a. in der unterschiedlichen Prognosedauer und den eingesetzten Verfahren zu suchen. Lediglich für DEMAT 3+ liegen bisher Befunde zum prognostischen Mehrwert bei zusätzlicher Berücksichtigung der Intelligenz mit $r_{adj} = .47$ und der Mathematiknote mit $r_{adj} = .49$ vor (Roick, Gölitz & Hasselhorn, 2004).

Prognose von Schulnoten

Die Vorhersage von Schulnoten durch bildungsstandardbasierte Testleistungen innerhalb der Grundschulzeit wurde bislang nur in der Studie von Hildebrandt und Watermann (2015) im Fach Deutsch untersucht. Zur Ermittlung des Testwerts in Deutsch in der vierten Klasse der Primarstufe wurde über mehrere sprachliche Leistungsdomänen (Lesen, Hören, Sprachgebrauch, Rechtschreibung) hin gemittelt. Der Zusammenhang mit zukünftigen Deutschnoten in der sechsten Klasse der Sekundarstufe (1 für *sehr gut* und 6 für *ungenügend*) betrug über alle Bildungsgänge hinweg im Mittel r = -.38. Auch bei Kontrolle zahlreicher Kovariaten (Deutsch- und Mathematiknote, Geschlecht, Intelligenz, Mathematiktestleistung, sozioökonomischer Familienhintergrund, Migrationshintergrund) konnte der bildungsstandardbasierte Test in Deutsch in der Primarstufe die Deutschnote in der Sekundarstufe vorhersagen: Die bildungsgangspezifischen standardisierten Regressionsgewichte lagen zwischen β_{adj} = -.19 und β_{adj} = -.17.

Darüber hinaus gibt es nur Querschnittsanalysen zum Zusammenhang zwischen bildungsstandardbasierten Testleistungen und Schulnoten innerhalb der Primarstufe. Hier liegen die Korrelationen zwischen $-.72 \leq r \leq -.48$ in Mathematik und $-.63 \leq r \leq -.53$ in Deutsch.

Die Prognosekraft von kommerziell erhältlichen, standardisierten Testverfahren für zukünftige Schulnoten, die bis zu 2 Jahre später erzielt wurden, beträgt in Mathematik (mit dem DEMAT; Krajewski, Liehm & Schneider, 2004; Roick et al., 2004) zwischen r = -.64 und r = -.69. Die Studie von Roick et al. (2004) wies zudem einen prognostischen Mehrwert der Mathematikleistung bei Kontrolle der Mathematiknote von r_{adj} = -.46 nach. Für Deutsch (mit dem HASE; Roos, Schöler & Treutlein, 2007) liegt die Prognosekraft zwischen r = -.38 bis r = -.45 für spätere Deutsch-, Lese- und Rechtschreibnoten.

Prognose der Gymnasialempfehlung

Befunde zur Vorhersage der zukünftigen Übertritts- bzw. Gymnasialempfehlung liegen unseres Wissens für bildungsstandardbasierte Tests nicht vor, da es hierzu bislang nur Querschnittsstudien gibt. Für bildungsstandardbasierte Tests berichten Köller, Eßel-Ullmann und Paasch (2012) innerhalb der fünften Klasse einen positiven Zusammenhang mit dem realisierten Gymnasialbesuch mit Korrelationswerten von r = .53 und r = .55.

Testwerte im DEMAT können die Übergangsempfehlung substanziell vorhersagen und besitzen prognostischen Mehrwert über die Deutsch- und Mathematiknote hinaus. Für Leistungstests in Deutsch liegen keine entsprechenden Befunde vor.

Forschungsfragen

Bildungsstandardbasierte Tests sind seit über 10 Jahren fester Bestandteil der deutschen Bildungspolitik und -forschung sowie der Schulpraxis: Sie gehören damit zu den am meisten eingesetzten standardisierten Tests an Schulen. Die Bildungsstandards implizieren, dass die Ergebnisse bildungsstandardbasierter Tests auch die zukünftige schulische Entwicklung von Schülerinnen und Schülern vorhersagen können. Jedoch gibt es bislang kaum Studien, die diese Fragestellung empirisch untersucht haben. In der vorliegenden Arbeit analysieren wir daher auf Grundlage repräsentativer Daten von Brandenburger Grundschulkindern die Vorhersagekraft von bildungsstandardbasierten Tests im Hinblick auf drei Forschungsfragen: Wie gut können bildungsstandardbasierte Tests (1) im Fach Mathematik und (2) im Fach Deutsch spätere Testleistungen und Schulnoten desselben Faches vorhersagen? (3) Wie hoch ist die gemeinsame Vorhersagekraft von Testleistungen in Mathematik und Deutsch für die Gymnasialempfehlung? Bei der Beantwortung dieser Forschungsfragen untersuchen wir nicht nur bivariate Zusammenhänge im Längsschnitt, sondern auch den prognostischen Mehrwert von bildungsstandardbasierten Kompetenztests bei Kontrolle von Schulnoten, Intelligenz und sozioökonomischem Hintergrund der Kinder.

Welche Ergebnisse sind zu erwarten? Inhaltlich relevante Kompetenzen bzw. das Vorwissen zu einem früheren Zeitpunkt gehören zu den besten Prädiktoren für den zukünftigen schulischen Kompetenz- und Wissenserwerb (Helmke & Weinert, 1997). Wie stark die längsschnittlichen Zusammenhänge tatsächlich sind, variiert (wie unser Literaturüberblick bzw. Tabelle ESM-1 zeigen) jedoch zwischen den eingesetzten Prädiktoren, den untersuchten Kriterien und der Prognosedauer. Auch wenn bislang für den Grundschulbereich in Deutschland nur wenige Befunde zur Verfügung stehen, stellen diese die beste Basis für eine Schätzung der zu erwartenden Ergebnisse in der vorliegenden Studie dar. Dabei können Befunde aus Querschnittsstudien (z. B. zu Schulnoten) als Obergrenze der zu erwartenden Vorhersagekraft im Rahmen dieser Längsschnittstudie angenommen werden. Schmal ist auch die Befundlage zum prognostischen Mehrwert bildungsstandardbasierter Tests über weitere leistungsprädiktive Merkmale. Und es ist darüber hinaus schwierig, Erwartungen zu

formulieren, da der Kranz an Kovariaten zwischen den Studien variiert. Generell sind für die vorliegende Studie Koeffizienten im unteren Wertebereich zu erwarten, da wir im Vergleich zu bisherigen Studien mehr leistungsprädiktive Merkmale und einen längeren Prognosezeitraum berücksichtigen.

Methode

Stichprobe und Prozedur

Zur Analyse der Forschungsfragen nutzten wir die Daten aus der Längsschnittstudie KEGS (Kompetenzentwicklung in der Grundschule), die an öffentlichen Grundschulen in Brandenburg im Auftrag des Ministeriums für Bildung, Jugend und Sport durchgeführt wurde (Fuchs & Brunner, 2014). Im Zentrum der KEGS-Studie stand die Kompetenzentwicklung von Grundschulkindern in Mathematik und Deutsch (Lesekompetenz). Die Erhebungen begannen im Schuljahr 2006/07 (zweite Klasse) und endeten im Schuljahr 2010/11 (sechste Klasse); sie wurden jährlich jeweils gegen Ende eines Schuljahres in den Monaten Mai bis Juli durchgeführt. Die mathematische Kompetenz wurde von der zweiten bis sechsten Klasse, die Lesekompetenz von der vierten bis zur sechsten Klasse erfasst. Abbildung 1 stellt die Anlage der KEGS-Studie dar.

Zu Beginn der KEGS-Studie war es den Schulleitungen freigestellt, ob sie teilnahmen oder nicht. Sofern eine Schulleitung zustimmte, waren alle Kinder einer zufällig gezogenen Klasse in der zweiten Jahrgangsstufe zur Teilnahme verpflichtet. Von anfänglich 100 zufällig gezogenen Schulen nahmen in der zweiten Klasse letztlich Kinder von 77 Schulen (77 %) teil; in der dritten Klasse waren es 75 Schulen. Um die Integrität der Stichprobe zu sichern, wurden ab dem Schuljahr 2008/09 (vierte Jahrgangsstufe) die Schulleitungen zur Teilnahme verpflichtet, wobei zugesichert wurde, dass die Daten auf Individual- und Schulebene allenfalls in anonymisierter Form an das Bildungsministerium weitergereicht werden. Dennoch reduzierte sich die Anzahl der teilnehmenden Schulen bis zum Zeitpunkt der sechsten Klasse, je nach untersuchtem Fach, auf 68 Schulen für Mathematik und auf 59 Schulen für Deutsch. Analysen zum Stichprobenausfall (s. ESM-2) zeigten, dass Kinder an Schulen, die weiterhin an KEGS teilnahmen, bei 9 von 10 Leistungsindikatoren (Kompetenztests, Noten, Intelligenz) besser abschnitten; Unterschiede in soziodemografischen Merkmalen waren nicht systematisch.

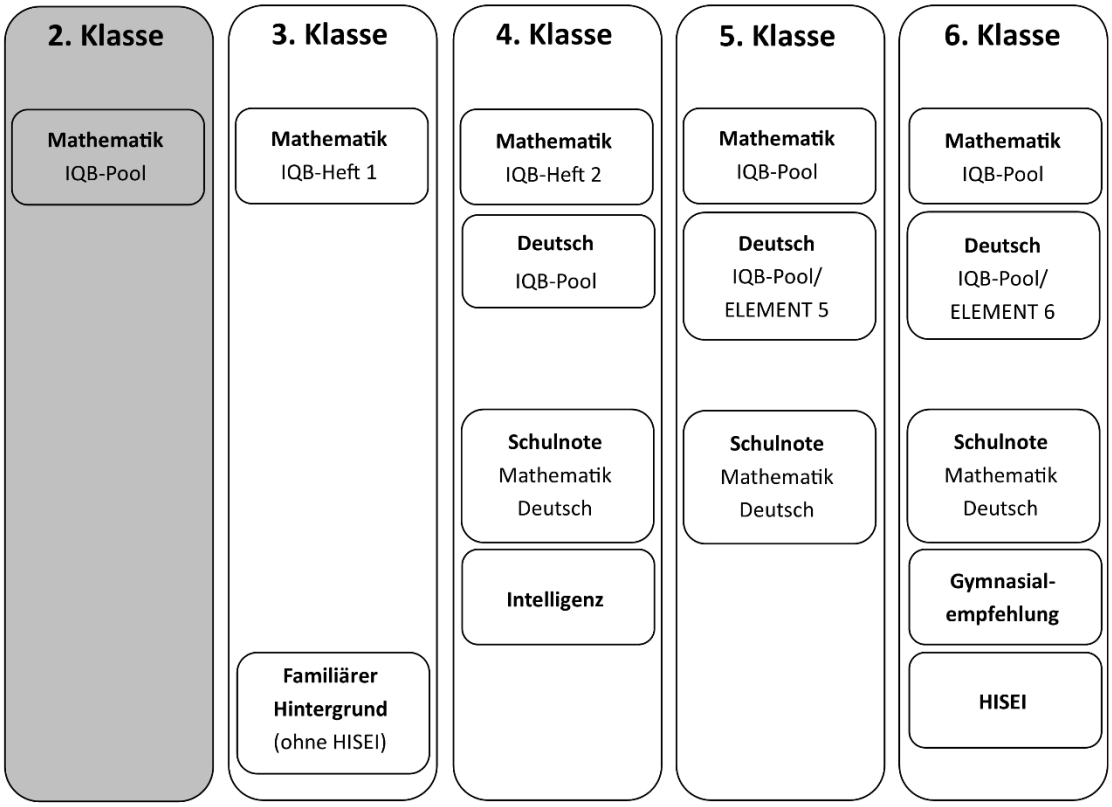


Abbildung 1. Anlage der KEGS-Studie über die Erhebungszeitpunkte von der 2. bis zur 6. Klasse. Die Daten in der 2. Klasse wurden nicht in die Analysemodelle dieses Artikels einbezogen (grau hinterlegt). IQB = Institut zur Qualitätsentwicklung im Bildungswesen.

Das Studiendesign von KEGS wurde nach Beginn stetig erweitert, um eine Vielzahl von Forschungsfragen zu beantworten, unter anderen auch die Frage nach Durchführungseffekten auf die Datenqualität von bildungsstandardbasierten Tests. Daher führten in einigen Schuljahren entweder Lehrkräfte oder geschulte Testleiterinnen und -leiter die Tests durch. Die Art der Testdurchführung wurde per Zufall festgelegt; in beiden Fällen bewerteten geschulte Kodiererinnen und Kodierer die Testantworten. Um die Vorhersagegüte von bildungsstandardbasierten Tests bestmöglich abschätzen zu können, fokussieren wir in den vorliegenden Analysen auf Daten, die auf einer Testadministration und -durchführung durch geschulte Testleiterinnen und -leiter basieren. (ESM-7 dokumentiert die Ergebnisse, die bei einer Durchführung der Mathematiktests durch Lehrkräfte in der dritten Klasse resultierten; s. a. Abschnitt 4.4). Da in Mathematik die Tests in der zweiten Klasse ausschließlich von Lehrkräften durchgeführt wurden, wurden Daten für Mathematik ab der dritten Klasse für die Analysen zur Vorhersagekraft berücksichtigt. Die so resultierende Längsschnittstichprobe umfasste 568 Kinder (49 % Mädchen; Alter in der dritten Klasse: M = 9.5 Jahre, SD = 0.6 Jahre) jeweils einer Klasse aus 22 Schulen (durchschnittlich 26 Kinder/Klasse), die an

mindestens einer Erhebung von der dritten bis zur sechsten Jahrgangsstufe teilnahmen (s. a. Abschnitt 3.3.2 zu fehlenden Werten).

Messinstrumente

Kompetenzen in Mathematik und Deutsch (Lesen)

Bei der Erfassung der Kompetenz in Mathematik wurden über alle Erhebungszeitpunkte vier von insgesamt fünf inhaltsbezogenen Kompetenzen berücksichtigt: Zahlen und Operationen, Muster und Strukturen, Größen und Messen sowie Raum und Form. Weiterhin wurden Tests zur Erfassung der Lesekompetenz im Fach Deutsch eingesetzt (Bremerich-Vos & Böhme, 2009). Die vorgesehene Bearbeitungsdauer der Kompetenztests in Mathematik und Deutsch betrug jeweils maximal 45 Minuten. Die Kinder bearbeiteten die Kompetenztests an separaten Tagen, mit maximal einer Woche Abstand. Die Schulen konnten wählen, welchen Test sie zuerst bearbeiteten. Die Items der Kompetenztests beider Fächer stammten zur Mehrzahl aus einem Aufgabenpool, der vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) zur Überprüfung der Bildungsstandards in der Primarstufe entwickelt wurde: In Mathematik wurde konkret in der dritten Klasse Heft 1, in der vierten Klasse Heft 2 eingesetzt (Granzer et al., 2008a, 2008b). Die Bildungsstandards fokussieren auf die vierte Jahrgangsstufe, und entsprechende Items sind auf das Leistungsniveau in dieser Jahrgangsstufe abgestimmt. Um die Kompetenzentwicklung in Deutsch in den Jahrgangsstufen 5 und 6 zu untersuchen, wurden zudem Items aus der ELEMENT-Studie eingesetzt (Lehmann & Lenkeit, 2008). Die Reliabilität (interne Konsistenz, KR-20) der Kompetenztests variierte zwischen den Erhebungszeitpunkten. In Mathematik lag die Reliabilität bei r_{tt} = .83, .80, .87 und .91 in der 3., 4., 5. und 6. Klassenstufe. Für den Kompetenztest in Deutsch lag die Reliabilität bei r_{tt} = .72, .74 und .81 in der 4., 5. und 6. Klassenstufe. Diese Werte sind für die vorliegenden Forschungsfragen als zumindest zufriedenstellend zu bewerten (Kline, 2005) und mit den Angaben von Köller, Eßel-Ullmann und Paasch (2012) vergleichbar.

Schulnoten und Gymnasialempfehlung

Im Rahmen der Erhebungen in der vierten, fünften und sechsten Klasse gaben die Lehrkräfte die Halbjahresnoten der Kinder in den Fächern Deutsch und Mathematik auf der Schulnotenskala von 1 (*sehr gut*) bis 6 (*ungenügend*) an. In der sechsten Jahrgangsstufe nannten die Lehrkräfte für alle Kinder ihrer Klasse die Empfehlung für einen Bildungsgang in

der Sekundarstufe I. In Brandenburg gab es zum Zeitpunkt der KEGS-Studie folgende Optionen: erweiterte Berufsbildungsreife (erweiterter Hauptschulabschluss), Fachoberschulreife (Mittlerer Schulabschluss) und Abitur (allgemeine Hochschulreife). Für die nachfolgenden Analysen kodierten wir die Gymnasialempfehlung (=Abitur) mit 1, alle anderen Optionen wurden mit 0 kodiert.

Intelligenz

In der vierten Klasse wurde der Untertest Figurenanalogien (N2, Testform A) des Kognitiven Fähigkeitstests (KFT 4-12+R; Heller & Perleth, 2000) durchgeführt. Figurenanalogien gelten als guter Indikator für fluide Intelligenz (Horn & Noll, 1997) und haben sich in zahlreichen Schulleistungsstudien bewährt. Die interne Konsistenz bei Kindern in der vierten Klasse liegt bei .94 (Heller & Perleth, 2000).

Familiärer Hintergrund

Informationen zum familiären Hintergrund stammten von den Kindern zum Zeitpunkt der sechsten Klasse sowie von den Eltern, die einen Elternfragebogen ausfüllten, als ihre Kinder die dritte Klasse besuchten. Die Kinder machten Angaben zum Beruf der Eltern. Diese Angaben wurden in den International Socio-Economic Index of Occupational Status (ISEI; Ganzeboom, De Graaf & Treiman, 1992) transformiert. Dabei wurde der höchste ISEI-Wert einer Familie, HISEI, für die vorliegenden Analysen herangezogen. Die Eltern gaben ihren schulischen und beruflichen Abschluss an; für die Analysen verwendeten wir den höchsten Abschluss in einer Familie. Weiterhin beantworteten die Eltern Fragen zum Buchbesitz, zur Häufigkeit gemeinsamer kultureller Aktivitäten, zum Vorhandensein von Wohlstandsgütern und zum monatlich zur Verfügung stehenden Nettoeinkommen des Haushaltes (zu Detailinformationen der Erfassung s. ESM-3).

Datenanalyse

Skalierung der Kompetenztests

Die Skalierung der Kompetenztests erfolgte für Mathematik und Lesen jeweils getrennt mit dem Programm ConQuest (Wu, Adams & Wilson, 1998) auf Basis eines eindimensionalen Rasch-Modells (Rasch, 1980) für dichotome Daten. Die Itemanalysen erfolgten in zwei

Schritten. Erstens wurden in enger Anlehnung an die methodischen Standards der PISA-Studie Items ausgeschlossen, die keine ausreichende Rasch-Homogenität aufwiesen (Organisation for Economic Co-operation and Development, 2009) und deren Trennschärfen nicht substanziell waren. Zweitens haben wir für die Skalierung der Testwerte in Mathematik und Lesen die Itemparameter aus den Normierungsstudien des IQB zugrunde gelegt. Dabei analysierten wir die Messinvarianz über die jeweiligen Erhebungszeitpunkte (Details zur Verlinkung s. ESM-4), indem wir das methodische Vorgehen zur Analyse aus der TOSCA-Studie (Nagy & Neumann, 2010) übernommen und die Klassifikation zum Differential-Item-Functioning (DIF) des Educational Testing Service angewendet haben. In den Analysen wurden dann nur Items berücksichtigt, die nicht der DIF-Kategorie C ($\beta_2 \leq 0.4$) angehörten. Letztendlich lagen so pro Erhebungszeitpunkt mindestens 21 Items zur Schätzung der Leistungskennwerte vor (im Detail, s. Fuchs & Brunner, 2014). Als Kompetenztestwerte ermittelten wir *Warms Weighted Maximum Likelihood Estimators* (WLE; Warm, 1989), die für die nachfolgenden Analysen in die KMK-Bildungsstandardmetrik (Bista-Metrik) transformiert wurden (Mittelwert für Gesamtdeutschland $M = 500$, $SD = 100$).

Umgang mit fehlenden Werten

Der Anteil fehlender Werte lag bei den Kompetenztestwerten in Mathematik und Deutsch zwischen 15 und 29 % und bei den übrigen Variablen zwischen 16 und 57 % (s. ESM-5). Wir nutzten das multiple Imputationsverfahren MICE (van Buuren & Groothuis-Oudshoorn, 2011), um 15 vollständige Datensätze zu erzeugen. Bei der Imputation der fehlenden Werte berücksichtigten wir das Skalenniveau der einbezogenen Variablen sowie die hierarchische Datenstruktur. Um die Qualität der imputierten Daten zu verbessern, nutzten wir zusätzlich zu den Variablen aus den Analysemodellen Alter, Geschlecht und die Mathematikleistung in der zweiten Klasse als Hilfsvariablen (Collins, Schafer & Kam, 2001).

Statistische Analysemodelle

Zur Analyse der Forschungsfragen 1 und 2 zur Vorhersagekraft bildungsstandardbasierter Tests in Mathematik und Deutsch für Testleistungen und Schulnoten verwendeten wir bivariate lineare Regressionsmodelle, in denen jeweils die Testleistung in einer bestimmten Jahrgangsstufe als Prädiktor und die Testleistung bzw. Schulnote in einer späteren Jahrgangsstufe als Kriterium spezifiziert wurden. Diese werden in den Tabellen 1 und 2

jeweils mit *Modell a* bezeichnet. Neben dem standardisierten Regressionskoeffizienten greifen wir zur Evaluation der gefundenen Zusammenhänge auch auf die Bildungsstandard-Metrik zurück. Ein Leistungsunterschied von 70 Punkten entspricht dabei der Breite einer Kompetenzstufe (Kultusministerkonferenz, 2013a, 2013b). Demnach geben die mit Faktor 70 multiplizierten unstandardisierten Regressionskoeffizienten (*B**70) an, wie sehr sich eine spätere Testleistung bzw. eine Schulnote verbessert, falls sich die Leistung auf der Prädiktorvariable um eine Kompetenzstufe erhöht[1].

Um den prognostischen Mehrwert bildungsstandardbasierter Tests abzuschätzen, analysierten wir multiple lineare Regressionsmodelle (*Modelle b*, Tabelle 1 und 2), in denen wir die bivariaten Modelle um zusätzliche Kontrollvariablen erweiterten: Hierzu gehörten die Schulnote (des korrespondierenden Faches; s. aber auch die Anmerkungen Tabelle 1), die Intelligenz und alle Indikatoren des familiären Hintergrundes. Zur Bewertung des prognostischen Mehrwerts berichten wir den quadrierten Semipartialkorrelationskoeffizienten (sr^2; Berechnung nach Cohen, Cohen, West & Aiken, 2003, S. 84ff.): Er gibt an, wie hoch der Anteil der Varianzaufklärung ist, welcher (unter Berücksichtigung aller Kontrollvariablen) allein auf den bildungsstandardbasierten Test zurückgeführt werden kann.

Zur Ermittlung der Vorhersagekraft bildungsstandardbasierter Tests für die Gymnasialempfehlung in der sechsten Klasse (Forschungsfrage 3) nutzten wir multivariate logistische Regressionsmodelle. Da die Lehrkräfte die Übertrittsempfehlung auf Grundlage der Leistungen in mehreren Fächern geben sollen (§15 Abs.1 GV, Ministerium für Bildung, Jugend und Sport, 2007), nahmen wir in einem ersten Analyseschritt die Testleistungen in Mathematik und Deutsch der vierten Klasse auf. Wir wählten hier die vierte Klasse, da in dieser Klassenstufe die Kinder zum ersten Mal Kompetenztests in beiden Fächern bearbeiteten. Im zweiten Analyseschritt untersuchten wir den prognostischen Mehrwert der Kompetenztests bei Kontrolle weiterer Kovariaten: Schulnoten desselben Fachs in der vierten Klasse, Intelligenz und Indikatoren des familiären Hintergrundes.

[1] Das methodische Vorgehen zur Sicherung der Messinvarianz hatte zur Folge, dass die Anzahl an sogenannten „Linking"-Items zur Verbindung der Tests von aufeinanderfolgenden Erhebungszeitpunkten in Mathematik zwischen 6 bis 21 Items und in Deutsch zwischen 6 bis 22 Items lag (s. OS-4). Bisher existieren keine einheitlichen Richtlinien, welche Anzahl an Items für eine valide Verlinkung notwendig ist (u. a. Kolen & Brennan, 2004). Eventuell bestehende Schwächen in der Metrikanbindung wirken sich in den Analysen zur Prognosekraft jedoch nur auf die unstandardisierten Regressionskoeffizienten (*B**70; s. Tabelle 1-3) aus, da diese Koeffizienten sich auf die Bista-Metrik beziehen. Die unstandardisierten Regressionskoeffizienten sollten daher mit gewisser Vorsicht interpretiert werden. Diese Einschränkung gilt jedoch nicht für die standardisierten Regressionskoeffizienten (β), die sich auf Standardabweichungseinheiten beziehen und somit unabhängig von der Metrik der unabhängigen und abhängigen Variablen sind.

Alle Modelle wurden mit M*plus* 7.3 (Muthén & Muthén, 1998-2012) berechnet. Zur Schätzung aller Modellparameter verwendeten wir das *Robust-Maximum-Likelihood*-Verfahren (MLR). Zudem berücksichtigten wir die hierarchische Datenstruktur unter Verwendung des Moduls „complex“ und aggregierten die Ergebnisse über 15 Datensätze.

Ergebnisse

Forschungsfrage 1: Vorhersagekraft des Kompetenztests in Mathematik

Die Leistungen der Kinder, die sie bei Kompetenztests in Mathematik in der dritten bzw. vierten Klasse zeigten, konnten Testleistungen in Mathematik, die sie bis zu 3 Jahre später erzielten, vorhersagen: Die Vorhersagegüte lag hier zwischen $\beta = .56$ und $\beta = .66$ (Tabelle 1, Testleistung, Modell a). Ein Leistungsunterschied von einer Kompetenzstufe war dabei mit 38 bis 50 Punkten besseren Leistungen in Mathematik in späteren Schuljahren assoziiert (Zeile *B**70 in Tab. 1, Modell a). Für alle Vorhersageintervalle zeigte sich ein deutlicher prognostischer Mehrwert der Mathematiktests zur Vorhersage der Testleistung mit adjustierten standardisierten Regressionskoeffizienten von $\beta_{adj} = .30$ bis $\beta_{adj} = .45$ (Tab. 1, Modell b). Selbst bei Kontrolle der Intelligenz, der Mathematiknote sowie den Merkmalen des familiären Hintergrundes lag die für einen Leistungsunterschied von einer Kompetenzstufe erwartete zukünftige Leistungsdifferenz zwischen 20 und 34 Punkten auf der Bista-Metrik. Die quadrierten Semipartialkorrelationskoeffizienten lagen zwischen $sr^2 = .06$ und $sr^2 = .11$, d. h. durch Kenntnis der Testleistung in Mathematik in der dritten bzw. vierten Jahrgangsstufe konnten (bei Kontrolle der Kovariaten) 6 bis 11 % mehr Varianz in späteren Testleistungen erklärt werden.

Ein ähnliches Befundmuster zeigte sich auch für die Vorhersage der Mathematiknoten. Kinder mit besseren Testleistungen in Mathematik erzielten in höheren Klassenstufen bessere Mathematiknoten. Die standardisierten Regressionskoeffizienten lagen zwischen $\beta = -.51$ und $\beta = -.57$ (Tab.1, Modell a). Leistungsunterschiede von einer Kompetenzstufe waren dabei unabhängig vom Prognosezeitraum mit besseren Mathematiknoten von etwa einem Drittel Notenpunkt assoziiert. Auch bei Kontrolle der Kovariaten für die Vorhersage der Mathematiknoten wiesen die Kompetenztests in Mathematik einen statistisch bedeutsamen, prognostischen Mehrwert auf (Tab. 1, Modell b). Die Werte lagen zur Vorhersage der Noten in der fünften und sechsten Klasse bei $\beta_{adj} = -.12$ und $\beta_{adj} = -.22$. Ein Leistungsunterschied von einer Kompetenzstufe war dabei mit -.07 und -.14 besseren Mathematiknoten assoziiert.

Die quadrierten Semipartialkorrelationskoeffizienten lagen zwischen $sr^2 = .01$ und $sr^2 = .03$. Zur Vorhersage der Noten in der vierten Klasse (ohne Kontrolle der Mathematiknote in der dritten Klasse) lagen die Werte erwartungsgemäß höher.

Tabelle 1

Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik

		Vorhersage mit Mathematik-Kompetenztest in der 3. Klasse auf Kriterien in der						Vorhersage mit Mathematik-Kompetenztest in der 4. Klasse auf Kriterien in der			
Modell		4. Klasse		5. Klasse		6. Klasse		5. Klasse		6. Klasse	
Kriterium: Testleistung in Mathematik											
(a)	β	.56	[.49, .64]	.63	[.57, .69]	.58	[.50, .66]	.66	[.59, .72]	.61	[.53, .70]
	*B**70	38	[33, 44]	47	[42, 52]	45	[38, 52]	50	[43, 58]	48	[41, 56]
	R^2	.32		.40		.34		.43		.38	
(b)	β_{adj}	.30[a]	[.21, .39][a]	.40[a]	[.32, .49][a]	.33[a]	[.26, .46][a]	.45	[.35, .54]	.36	[.27, .46]
	B_{adj}*70	20[a]	[14, 27][a]	30[a]	[23, 37][a]	26[a]	[18, 33][a]	34	[26, 42]	29	[21, 36]
	R^2	.49[a]		.57[a]		.52[a]		.57		.53	
	sr^2	.06[a]		.11[a]		.07[a]		.11		.08	
Kriterium: Mathematiknote											
(a)	β	-.54	[-.63, -.45]	-.48	[-.56, -.39]	-.51	[-.60, -.41]	-.57	[-.66, -.48]	-.57	[-.66, -.49]
	*B**70	-.28	[-.33, -.23]	-.28	[-.33, -.23]	-.35	[-.49, -.21]	-.35	[-.42, -.28]	-.42	[-.49, -.35]
	R^2	.29		.23		.26		.32		.33	
(b)	β_{adj}	-.41[b]	[-.51, -.30][b]	-.12[a]	[-.21, -.03][a]	-.15[a]	[-.25, -.05][a]	-.22	[-.32, -.11]	-.22	[-.31, -.13]
	B_{adj}*70	-.21[b]	[-.26, -.16][b]	-.07[a]	[-.12, -.02][a]	-.07[a]	[-.12, -.02][a]	-.14	[-.28, .00]	-.14	[-.19, .09]
	R^2	.41[b]		.56[a]		.56[a]		.58		.58	
	sr^2	.14[b]		.01[a]		.01[a]		.03		.03	

Anmerkungen. Modell a: bivariates Regressionsmodell. Modell b: multiples Regressionsmodell (Spezifikation s. Text). In Klammern wird das 95%-Konfidenzintervall ausgewiesen. *B**70, B_{adj}*70 Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. R^2 = Determinationskoeffizient. sr^2 = Quadrierter Semipartialkorrelationskoeffizient.

[a] In der 3. Klasse wurden die Schulnoten nicht erfasst; in Modell b wurde daher für die Mathematiknote der 4. Klasse kontrolliert. [b] In der 3. Klasse wurden die Schulnoten nicht erfasst; in Modell b wurde daher nicht für die Mathematiknote kontrolliert.

Forschungsfrage 2: Vorhersagekraft des Lesekompetenztests in Deutsch

Die Lesekompetenzleistungen der Kinder in der vierten Klasse konnten – über Zeitintervalle von bis zu 2 Jahren – spätere Testleistungen in Deutsch vorhersagen: Die Vorhersagegüte lag bei $\beta = .60$ und $\beta = .61$ (Tab. 2, Modell a). Leistungsunterschiede von einer Kompetenzstufe waren dabei mit 36 und 42 Punkten besseren Leistungen in Deutsch in späteren Schuljahren assoziiert (Zeile *B**70, Tab. 2, Modell a). Darüber hinaus zeigte sich selbst bei Kontrolle der Kovariaten in Deutsch ein deutlicher prognostischer Mehrwert zur Vorhersage der Testleistung mit Werten von $\beta_{adj} = .43$ und $\beta_{adj} = .44$ (Tab. 2, Modell b). Hier lag die für einen Leistungsunterschied von einer Kompetenzstufe erwartete zukünftige Leistungsdifferenz bei 27 und 30 Punkten auf der Bista-Metrik. Die quadrierten Semipartialkorrelationskoeffizienten lagen jeweils bei $sr^2 = .13$.

Für die Vorhersage der Deutschnoten wurden standardisierte Regressionskoeffizienten von $\beta = -.47$ und $\beta = -.50$ ermittelt (Tab. 2, Modell a). Kinder mit besseren Testleistungen in Deutsch in der vierten Klasse erzielten in höheren Klassenstufen bessere Deutschnoten. Ein Leistungsunterschied von einer Kompetenzstufe war dabei mit -.21 und -.28 besseren Deutschnoten assoziiert. Für die Deutschtests zeigte sich darüber hinaus wiederum bei Kontrolle der Kovariaten ein prognostischer Mehrwert von $\beta_{adj} = -.19$ (Tab. 2, Schulnote, Modell b). So lag für einen Leistungsunterschied von einer Kompetenzstufe die erwartete zukünftige Leistungsdifferenz bei -.07 Notenpunkten. Die quadrierten Semipartialkorrelationskoeffizienten lagen bei $sr^2 = .01$ und $sr^2 = .02$.

Tabelle 2

Forschungsfrage 2: Vorhersage zukünftiger Testleistungen und Schulnoten in Deutsch

		Vorhersage mit Deutsch-Lesekompetenztest in der 4. Klasse auf Kriterien in der			
Modell		5. Klasse		6. Klasse	
Kriterium: Testleistung in Deutsch-Lesen					
(a)	β	.60	[.53, .67]	.61	[.54, .68]
	*B**70	36	[31, 42]	42	[37, 48]
	R^2	.36		.37	
(b)	β_{adj}	.44	[.35, .53]	.43	[.34, .52]
	B_{adj}*70	27	[20, 33]	30	[24, 36]
	R^2	.45		.51	
	sr^2	.13		.13	
Kriterium: Deutschnote					
(a)	β	-.47	[-.55, -.38]	-.50	[-.58, -.42]
	*B**70	-.21	[-.25, -.17]	-.28	[-.33, -.23]
	R^2	.22		.25	
(b)	β_{adj}	-.19	[-.22, -.06]	-.19	[-.28, -.10]
	B_{adj} *70	-.07	[-.11, -.03]	-.07	[-.11, -.03]
	R^2	.58		.57	
	sr^2	.01		.02	

Anmerkungen. Modell a: bivariates Regressionsmodell. Modell b: multiples Regressionsmodell (Spezifikation s. Text). In Klammern wird das 95%-Konfidenzintervall ausgewiesen. *B**70, B_{adj}*70 Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. R^2 = Determinationskoeffizient. sr^2 = Quadrierter Semipartialkorrelationskoeffizient.

Forschungsfrage 3: Vorhersage der Gymnasialempfehlung

Bildungsstandardbasierte Kompetenztests in Mathematik und der Lesekompetenz in Deutsch in der vierten Klasse konnten die Gymnasialempfehlung in der sechsten Klasse vorhersagen (Tab. 3, Modell a). Die bivariaten (punktbiserialen) Korrelationen zwischen Gymnasialempfehlung in der sechsten Klasse und den Kompetenztestwerten in Mathematik und Deutsch in der vierten Klasse lagen bei r_{pb} = .68 und r_{pb} = .62 (s. ESM-5).

Die Analysen der logistischen Regressionsmodelle zeigten, dass höhere Kompetenzwerte in Mathematik bzw. der Lesekompetenz jeweils mit einer höheren Chance einhergingen, dass Lehrkräfte 2 Jahre später eine Gymnasialempfehlung aussprachen (bei gleichzeitiger Kontrolle der Kompetenz im jeweils anderen Fach; s. Tab. 3, Modell a). Die Regressionskoeffizienten können genutzt werden, um die Wahrscheinlichkeiten für den Erhalt

einer Gymnasialempfehlung zu veranschaulichen. Hierzu berechneten wir die adjustierten Wahrscheinlichkeiten für den Erhalt einer Gymnasialempfehlung an den Kompetenzstufengrenzen bei 390, 460, 530 und 600 Punkten auf der Bista-Metrik (Abbildung 2). Für die Kontrollvariablen wurde der jeweilige Stichprobenmittelwert einbezogen. Mit dem Zugewinn einer Kompetenzstufe in Mathematik ging eine höhere (für die Lesekompetenz adjustierte) Wahrscheinlichkeit der Gymnasialempfehlung von p = .14 bis .20 einher. Für Lesen lag die (für die Mathematikkompetenz adjustierte) Erhöhung der Wahrscheinlichkeit für eine Gymnasialempfehlung zwischen p = .11 und .14. Wichtig ist an dieser Stelle festzuhalten, dass sowohl der Kompetenztest in Mathematik als auch der Lesekompetenztest einen prognostischen Mehrwert zur Vorhersage der Gymnasialempfehlung aufwiesen (Tab. 3, Modell b). Bei zusätzlicher Berücksichtigung der Kovariaten lagen die Zuwächse der adjustierten Wahrscheinlichkeiten für eine Gymnasialempfehlung bei Werten im Kompetenztest Mathematik und im Lesekompetenztests zwischen p = .08 und .09.

Tabelle 3

Forschungsfrage 3: Vorhersage der Gymnasialempfehlung in der 6. Klasse auf Grundlage bildungsstandardbasierter Testleistungen in Mathematik und Deutsch in der 4. Klasse

Modell	Testdomäne	Koeffizient	95%-Konfidenzintervall
(a)	Mathematik		
	$Exp(\beta)_{adj}$	3.48	[2.45, 4.96]
	$Exp(B*70)_{adj}$	2.32	[1.76, 3.05]
	Deutsch		
	$Exp(\beta)_{adj}$	2.47	[1.80, 3.40]
	$Exp(B*70)_{adj}$	1.75	[1.53, 2.01]
(b)	Mathematik		
	$Exp(\beta)_{adj}$	1.78	[1.05, 3.02]
	$Exp(B*70)_{adj}$	1.42	[1.08, 1.87]
	Deutsch		
	$Exp(\beta)_{adj}$	1.81	[1.24, 2.65]
	$Exp(B*70)_{adj}$	1.42	[1.08, 1.87]

Anmerkungen. Modell a: multiples Regressionsmodell mit den Testleistungen in Deutsch und Mathematik als Prädiktoren (Spezifikation s. Text). Modell b: multiples Regressionsmodell, in dem zusätzlich zu den Testleistungen in Deutsch und Mathematik für zahlreiche leistungsrelevante Kovariaten kontrolliert wurde (Spezifikation s. Text). Exp(β) „Semistandardisierter" Wert, Vorhersagekraft für einen Leistungsunterschied von einer Standardabweichung des Prädiktors, Exp(*B**70) Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. Die Werte in der Spalte „Koeffizient" können auch als Odds Ratios interpretiert werden.

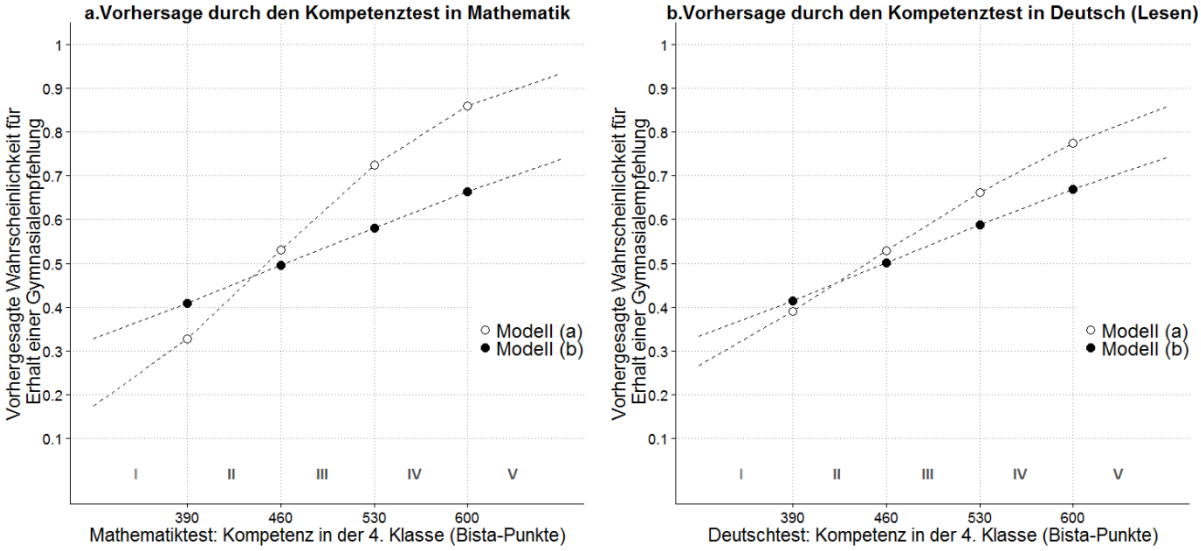


Abbildung 2. Forschungsfrage 3: Vorhergesagte Wahrscheinlichkeiten eine Gymnasialempfehlung am Ende der 6. Klasse zu erhalten auf Grundlage bildungsstandardbasierter Testleistungen (in Bista-Punkten) in (a) Mathematik und (b) Deutsch (Lesen). Modell a: multiples Regressionsmodell, lediglich für anderen Fachtest (Stichprobenmittelwert) kontrolliert (Spezifikation s. Text). Modell b: multiples Regressionsmodell, neben anderem Fachtest ergänzend für zahlreiche weitere Kovariaten (jeweils Stichprobenmittelwerte) kontrolliert (Spezifikation s. Text).

Zusatzanalysen

Im Rahmen von zwei Zusatzanalysen gingen wir der Frage nach, inwiefern die bisherigen Befunde repliziert und generalisiert werden können. Erstens kann durch das Zentrieren der Noten am jeweiligen Klassenmittelwert für etwaige Strenge- und Mildeeffekte, aber auch systematische Unterschiede im mittleren Notenniveau kontrolliert werden. Die Ergebnisse dieser Analysen sind im Online Supplement (s. ESM-6.1 bis 6.3) dokumentiert. Zweitens konnten wir potenzielle Effekte der Durchführungsmodalitäten untersuchen, da für eine zufällig bestimmte Teilstichprobe der KEGS-Studie Lehrkräfte (und nicht trainierte Testleiter/-innen) die Kompetenztests durchführten (N = 296 Kinder je einer Klasse mit durchschnittlich 23 Kindern aus 13 Schulen, 47 % Mädchen, Alter in der dritten Klasse: M = 9.5 Jahre, SD = 0.5 Jahre). Die Auswertung der Testantworten erfolgte auch hier durch trainierte Kodiererinnen und Kodierer. Einen Teil der Prognosemodelle (Vorhersagen ab der dritten Klasse für Mathematik) konnten wir mit diesem Teildatensatz der KEGS-Studie untersuchen (s. ESM-7.1 bis 7.4). Zusammenfassend ist festzuhalten, dass beide Zusatzanalysen das Ergebnismuster für die Prognosegüte replizierten, auch wenn es teilweise kleinere Unterschiede zwischen den Modellparametern gab: Die stärksten Abweichungen fanden wir für die adjustierten Regressionskoeffizienten zur Vorhersage der Schulnote in der sechsten Klasse in Abhängigkeit der Durchführungsmodalität (s. ESM-7.3). Trotz dieser kleineren Unterschiede untermauern die Zusatzanalysen die Generalisierbarkeit der oben darstellten Ergebnisse.

Diskussion

Prognosekraft bildungsstandardbasierter Tests

Im vorliegenden Beitrag untersuchten wir im Rahmen von drei Forschungsfragen die Prognosekraft bildungsstandardbasierter Tests in Mathematik und der Lesekompetenz in Deutsch für den schulischen Erfolg von Grundschulkindern. Im Rahmen der ersten beiden Forschungsfragen analysierten wir, wie gut bildungsstandardbasierte Tests (1) der Mathematikkompetenz und (2) der Lesekompetenz (in Deutsch) spätere Testleistungen und Schulnoten desselben Faches vorhersagen können. Es zeigte sich, dass Kompetenztests in Mathematik (für Prognosezeiträume von bis zu 3 Jahren) und Tests der Lesekompetenz in Deutsch (für Prognosezeiträume von bis zu 2 Jahren) spätere Testleistungen und Schulnoten desselben Faches substanziell vorhersagen können. Die standardisierten

Regressionskoeffizienten der bivariaten Regressionsmodelle zur Vorhersagegüte stimmen im Wesentlichen überein mit den wenigen relevanten Vorgängerstudien (z. B. Nachtigall, 2014, für Überblick zum Forschungsstand s. ESM-1). Die Vorhersagekraft für spätere Testleistungen und von Schulnoten war zudem vergleichbar mit der von kommerziell erhältlichen, standardisierten Schulleistungstests (z. B. DEMAT, HASE).

Ein prognostischer Mehrwert bildungsstandardbasierter Tests zur Erfassung der Kompetenz in Mathematik und der Lesekompetenz blieb auch dann bestehen, wenn für weitere leistungsprädiktive Merkmale, wie Schulnoten, Intelligenz und familiärer Hintergrund der Kinder in den Vorhersagemodellen kontrolliert wurde. Dies sind eindrucksvolle Befunde, da bislang (mit Ausnahme der Studie von Hildebrandt und Watermann, 2015) vergleichbar strenge Analysen zur Vorhersagekraft (bildungs-)standardisierter Leistungstests für den Primarbereich fehlten. Diese Ergebnisse stehen zudem auch im Einklang mit Befunden zur Sekundarstufe, in denen ein prognostischer Mehrwert von bildungsstandardbasierten Tests im Rahmen der Vergleichsarbeiten in der achten Jahrgangsstufe zur Vorhersage von Prüfungsnoten in der zehnten Klasse aufgezeigt werden konnte (Graf, Harych, Wendt, Emmrich & Brunner, in Druck). Generell helfen die vorliegenden Ergebnisse zum prognostischen Mehrwert, Kritik an schulischen Leistungstests zu entkräften, wonach diese lediglich den sozioökonomischen Status (Sackett et al., 2009) oder Intelligenz widerspiegeln (Baumert et al., 2009) sollen.

Die dritte Forschungsfrage befasste sich mit der Prognose der Gymnasialempfehlung. Die vorliegenden Ergebnisse zeigten, dass bildungsstandardbasierte Tests in Mathematik und Lesen eine bedeutsame, gemeinsame Vorhersagekraft für die Gymnasialempfehlung aufweisen. Die Vorhersagekraft für die Gymnasialempfehlung verminderte sich bei Kontrolle der Kovariaten (Schulnoten, Intelligenz und Indikatoren des familiären Hintergrunds) etwas, blieb aber dennoch statistisch bedeutsam.

Grenzen der Studie

Als Datengrundlage der dargestellten Analysen diente die KEGS-Längsschnittstudie, deren Studienanlage während des Verlaufs der Erhebungen stetig erweitert wurde. Bei der Analyse der Daten zur Vorhersagegüte musste für das vorliegende Papier die Analysestrategie auf die Verfügbarkeit der Daten abgestimmt werden, und es lagen bedauerlicherweise nicht für alle

Fächer und Zeitpunkte alle relevanten Variablen vor: z. B. Schulnoten in der dritten Klasse oder die Deutschleistung in der dritten Klasse.

Darüber hinaus ist unklar, inwiefern sich die Ergebnisse von Brandenburger Grundschulkindern im Hinblick auf andere Schülerinnen und Schüler in Deutschland generalisieren lassen. So findet der Wechsel von der Primar- in die Sekundarstufe in anderen Bundesländern bereits nach der vierten Jahrgangsstufe und damit 2 Jahre früher als in Brandenburg statt. Bisherige Analysen von Hildebrandt und Watermann (2015) sowie Nachtigall (2014) für Schülerstichproben mit Übergang in die Sekundarstufe I nach der vierten Jahrgangsstufe konnten jedoch ebenso einen substanziellen Zusammenhang zwischen bildungsstandardbasierten Testleistungen in der Primarstufe mit zukünftigen Testleistungen und Schulnoten in der Sekundarstufe aufzeigen.

Des Weiteren lassen sich die vorliegenden Ergebnisse nicht uneingeschränkt auf bildungsstandardbasierte Tests, wie diese im Rahmen von bspw. VERA 3 eingesetzt werden, übertragen. So wurden in unserer Studie in den Mathematiktests vier von fünf inhaltsbezogenen Kompetenzen (Leitideen) erfasst sowie die Durchführung und Kodierung der Tests in beiden Fächern geschulten Testleiterinnen und Testleitern bzw. Kodiererinnen und Kodierern übertragen. Im Rahmen von VERA 3 werden hingegen die Schülerinnen und Schüler in der Regel in zwei mathematischen Leitideen geprüft und die Testdurchführung und -auswertung von Lehrkräften übernommen. Die Studie von Graf, Emmrich, Harych und Brunner (2013) zeigte, dass Leistungsunterschiede bei bildungsstandardbasierten Tests in Mathematik im Rahmen von VERA 8 aus der Durchführung durch Testleiterinnen und Testleiter einerseits und durch Lehrkräfte andererseits resultieren können, jedoch nicht zwangsläufig resultieren müssen. Spoden, Fleischer und Leutner (2014) konnten für VERA-8-Tests feststellen, dass eine geringere Objektivität bei Auswertung der Testantworten durch Lehrkräfte mit einer höheren mittleren Schülerkompetenz einhergeht. Entsprechend den Erfahrungen im angloamerikanischen Raum, in denen viele Kompetenzmessungen im Schulbereich high-stakes Charakter haben (Amrein-Beardsley, Berliner & Rideau, 2010), ist zu erwarten, dass Lehrkräfte den ihnen zur Verfügung stehenden Spielraum zur Beeinflussung der Ergebnisse durch intendierte oder nicht intendierte Handlungen umso mehr nutzen, je stärker sie VERA als ein Kontrollinstrument (bspw. der Schuladministration) wahrnehmen. Inwiefern solche Verzerrungen zu Unter- als auch zu Überschätzungen der Vorhersagekraft von VERA-Tests führen, ist eine offene empirische Frage. Im Rahmen unserer Zusatzanalysen zeigten sich allenfalls kleine Unterschiede in der Prognosegüte zwischen

Tests, die von Lehrerinnen und Lehrern oder Testleiterinnen und Testleitern administriert und durchgeführt wurden (s. Abschnitt 4.4 und ESM-7.1 bis 7.4).

Abschließend muss für unsere Ergebnisse zu den Leistungstests in Deutsch (Lesekompetenz) einschränkend darauf hingewiesen werden, dass sich vor allem der Test zum Zeitpunkt der sechsten Jahrgangsstufe fast vollständig aus ELEMENT-Aufgaben zusammensetzte. Im Rahmen der ELEMENT-Studie 5 und 6 wurden die Testhefte aus Aufgaben verschiedener Schulleistungstests zusammengestellt, die nicht explizit nach den Bildungsstandards konzipiert wurden (Lehmann & Lenkeit, 2008).

Schlussfolgerung

Bildungsstandardbasierte Tests werden vor allem im Rahmen der Vergleichsarbeiten in pädagogischen Kreisen heftig kritisiert (u. a. Kuhn, 2014). Darüber hinaus konnten Nachtigall und Hellrung (2013) über die letzten 10 Jahre einen stagnierenden bis rückläufigen Trend bei der wahrgenommen Nützlichkeit von VERA feststellen. Der wahrgenommene Nutzen und die Akzeptanz von Ergebnisrückmeldungen sind aber zentrale Kriterien für die weitere Nutzung der im Rahmen von VERA gewonnenen Informationen (Bonsen, Büchter, & Peek, 2006; Kühle & Peek, 2007; Maier, 2008). Unsere Befunde zur Prognosegüte bildungsstandardbasierter Tests liefern Hinweise für die (praktische) Relevanz und Nützlichkeit solcher Tests.

So weisen die Befunde dieser Studie darauf hin, dass Lehrkräfte durch Ergebnisse aus bildungsstandardbasierten Tests wichtige Informationen zum zukünftigen Schulerfolg ihrer Schülerinnen und Schüler erhalten können. Wie die Befunde zum prognostischen Mehrwert zeigten, sind diese Informationen auch nicht durch das Wissen um Schulnoten, Intelligenztestergebnisse oder den familiären Hintergrund zu ersetzen. Das Potenzial von bildungsstandardbasierten Tests zur Kompetenzbeurteilung wird vor allem darin gesehen, dass die Schülerleistungen durch die Verortung auf den Bildungsstandards auf einem kriterialen Maßstab zurückgemeldet werden. So können Leistungen von Schülerinnen und Schülern auf einem klassenübergreifenden Referenzrahmen verortet werden, was vielen Lehrkräften nur bedingt gelingt (Brunner, Anders, Hachfeld & Krauss, 2011). Gerade deshalb werden Schulnoten auf ihre Vergleichbarkeit kritisiert (zusammenfassend Lintorf, 2012). Bildungsstandardbasierte Tests haben somit das Potenzial, Lehrkräften im Sinne eines *Screenings* eine Informationsgrundlage zu verschaffen, die die Erkenntnisse aus ihrer eigenen

Diagnostik im schulischen Alltag komplementiert und anreichert. In dieser Logik stellen bildungsstandardbasierte Tests einen Ausgangspunkt für weiterführende individuelle Feindiagnostik dar (Köller, Reiss, Stanat & Pant, 2012; Leutner et al., 2008).

Literatur

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Amrein-Beardsley, A., Berliner, D.C. & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, 18* (14), 1–36.

Baumert, J., Lüdtke, O., Trautwein, U. & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4* (3), 165–176.

Bonsen, M., Büchter, A. & Peek, R. (2006). Datengestützte Schul- und Unterrichtsentwicklung. Bewertung der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. In B. W., H.G. Holtappels, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Band 14, S. 125–148).

Bremerich-Vos, A. & Böhme, K. (2009). Lesekompetenzdiagnostik. Die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 214–249). Weinheim: Beltz.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Münster: Waxmann.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal Of Statistical Software, 45* (3), 1–67.

Collins, L.M., Schafer, J.L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures (Special Section: New Approaches to Missing Data.). *Psychological Methods, 6* (4), 330–351.

Fuchs, G. & Brunner, M. (Hrsg.). (2014). *Kompetenzentwicklung und Schulqualität an Brandenburger Grundschulen. Abschlussbericht der KEGS-Studie.* Berlin: Institut für Schulqualität der Länder Berlin und Brandenburg e.V. .

Ganzeboom, H.B.G., De Graaf, P.M. & Treiman, D.J. (1992). A Standard International Socio-Economic Index of Occupational Status. *Social Science Research*, *21* (1), 1–56.

Graf, T., Emmrich, R., Harych, P. & Brunner, M. (2013). Durchführungseffekte bei Vergleichsarbeiten in Jahrgangsstufe 8. *Empirische Pädagogik*, *27* (4), 459–473.

Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (in Druck). Wie gut können VERA-8 Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie*.

Granzer, D., Köller, O., Reiss, K., Robitzsch, A., Walther, G. & Winkelmann, H. (2008a). *Bildungsstandards. Kompetenzen überprüfen. Grundschule. Klasse 3/4 – Heft 1*. Berlin: Cornelsen Verlag.

Granzer, D., Köller, O., Reiss, K., Robitzsch, A., Walther, G. & Winkelmann, H. (2008b). *Bildungsstandards. Kompetenzen überprüfen. Grundschule. Klasse 3/4 – Heft 2*. Berlin: Cornelsen Verlag.

Heller, K.A. & Perleth, C. (2000). *KFT 4-12+R. kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* (3. Auflage). Göttingen: Beltz Test.

Helmke, A. & Weinert, F.E. (1997). Bedingungsfaktoren schulischer Leistungen. *Enzyklopädie der Psychologie: Psychologie des Unterrichts und der Schule* (S. 71–176). Göttingen: Hogrefe.

Hildebrandt, J. & Watermann, R. (2015, März). *Prognostische Validität von curricular gemessenen Testleistungen am Ende der Grundschulzeit*. Vortrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Bochum.

Horn, J.L. & Noll, J. (1997). Human cognitive capabilities: Gf–Gc theory. *Contemporary intellectual assessment : Theories, tests, and issues* (S. 53–91). New York: The Guilford Press.

Kline, R.B. (2005). *Principles and practice of structural equation modeling* (2. Auflage). New York: Guilford Press.

Kultusministerkonferenz (Hrsg.). (2004). *Bildungsstandards der Kultusministerkonferenz – Erläuterungen zur Konzeption und Entwicklung (Am 16.12.2004 von der Kultusministerkonferenz zustimmend zur Kenntnis genommen).*

Kultusministerkonferenz (Hrsg.). (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* Bonn.

Kultusministerkonferenz (Hrsg.). (2012). *Vereinbarung zur Weiterentwicklung von VERA. (Beschluss der Kultusministerkonferenz vom 08.03.2012).*

Kultusministerkonferenz (Hrsg.). (2013a). *Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Primarbereich (Jahrgangsstufe 4) (Auf Grundlage des Ländervergleichs 2011 überarbeitete Version in der Fassung vom 11. Februar 2013).*

Kultusministerkonferenz (Hrsg.). (2013b). *Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich „Lesen – mit Texten und Medien umgehen" – Primarbereich – (Auf Grundlage des Ländervergleichs 2011 überarbeiteter Entwurf in der Version vom 13. Februar 2013).*

Kultusministerkonferenz (Hrsg.). (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring: Beschluss der 350. Kultusministerkonferenz vom 11.06.2015*. Bonn.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling and linking. Methods and practices* (2. Auflage). New York: Springer.

Köller, O., Eβel-Ullmann, G. & Paasch, D. (2012). Validierung eines Instruments zur Erfassung Standard-basierter mathematischer Kompetenzen in der Grundschule. = Validation of an instrument on standard-based mathematical competencies in primary school. *Psychologie in Erziehung und Unterricht, 59* (3), 177–190.

Köller, O., Reiss, K., Stanat, P. & Pant, H.A. (2012). Diagnostik Standard-basierter mathematischer Kompetenzen im Primarbereich. Ein Überblick. *Psychologie in Erziehung und Unterricht, 59* (3), 163–176.

Krajewski, K., Liehm, S. & Schneider, W. (2004). *DEMAT 2+. Deutscher Mathematiktest für zweite Klassen*. Göttingen: Beltz.

Kühle, B. & Peek, R. (2007). Lernstandserhebungen in Nordrhein-Westfalen. Evaluationsbefunde zur Rezeption und zum Umgang mit Ergebnisrückmeldungen in Schulen. *Empirische Pädagogik, 21* (4), 428–447.

Kuhn, H.-J. (2014). Anspruch, Wirklichkeit und Perspektiven der Gesamtstrategie der KMK zum Bildungsmonitoring. *Die deutsche Schule, 106* (3), 414–426.

Lehmann, R. & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien.* Berlin: Humboldt Universität zu Berlin.

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2008). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin &

H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (S. 149–167). VS Verlag für Sozialwissenschaften.

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: Verlag für Sozialwissenschaften.

Maier, U. (2008). Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. *Zeitschrift für Erziehungswissenschaft, 11* (3), 453–474.

Ministerium für Bildung, Jugend und Sport. (2007). Verwaltungsvorschriften zur Grundschulverordnung (VV-GV).

Muthén, L.K. & Muthén, B.O. (1998–2012). *Mplus User's Guide*. Los Angeles: Muthén & Muthén.

Nachtigall, C. (2014). *Landesbericht. Thüringer Kompetenztests 2014*.

Nachtigall, C. & Hellrung, K. (2013). Zur zeitlichen Entwicklung der Rezeption von Vergleichsarbeiten. *Empirische Pädagogik, 27* (4), 423–441.

Nagy, G. & Neumann, M. (2010). Psychometrische Aspekte des Tests zu den voruniversitären Mathematikleistungen in TOSCA-2002 und TOSCA-2006: Unterrichtsvalidität, Rasch-Homogenität und Messäquivalenz. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten* (S. 281–306). VS Verlag für Sozialwissenschaften.

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 Technical Report*. Paris: OECD Publishing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Edition.). Chicago: University of Chicago Press.

Roick, T., Gölitz, D. & Hasselhorn, M. (2004). *DEMAT 3+. Deutscher Mathematiktest für dritte Klassen*. Göttingen: Beltz.

Roos, J., Schöler, H. & Treutlein, A. (2007). *Zur prognostischen Validität des Heidelberger Auditiven Screenings in der Einschulungsdiagnostik HASE. Abschlussbericht des Projektes EVER*. Heidelberg: Pädagogische Hochschule Heidelberg.

Sackett, P.R., Kuncel, N.R., Arneson, J.J., Cooper, S.R. & Waters, S.D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135* (1), 1–22.

Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik, 35* (1), 79–99.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54* (3), 427–450.

Wu, M.L., Adams, R.J. & Wilson, M. (1998). *ACER ConQuest*. Melbourne: The Australian Council for Educational Research Limited (ACER).

Elektronische Supplemente (ESM) – Studie I

Die elektronischen Supplemente sind mit der Onlineversion dieses Artikels verfügbar unter http://dx.doi.org/10.1024/1010-0652/a000195.

ESM-1.
Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Leistungstests.

ESM-2.
Überblick zum Stichprobenausfall der Schulen bis zur 6. Klasse für Mathematik (ab der 3. Klasse) und für Deutsch (ab der 4. Klasse).

ESM-3.
Darstellung der Items zur Erfassung des familiären Hintergrundes.

ESM-4.
Veranschaulichung der Messinvarianz „Verlinkung“ der bildungsstandardbasierten Kompetenztests in Mathematik und Deutsch (Lesen) mit Itemparametern aus dem IQB-Aufgabenpool und der ELEMENT-Studie in Klasse 5 und 6.

ESM-5.
Interkorrelationsmatrix aller Analysevariablen.

ESM-6.
Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik – Schulnoten klassenzentriert.

ESM-7.
Forschungsfrage 2: Vorhersage zukünftiger Testleistungen und Schulnoten in Deutsch – Schulnoten klassenzentriert.

ESM-8.
Forschungsfrage 3: Vorhersage der Gymnasialempfehlung in der 6. Klasse auf Grundlage bildungsstandardbasierter Testleistungen in Mathematik und Deutsch in der 4. Klasse – Schulnoten klassenzentriert.

ESM-9.

Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik – Testdurchführung von Lehrkräften (3. Klasse).

ESM-10.

Vergleichende Darstellung der standardisierten Regressionskoeffizienten (β) mit 95 %-Konfidenzintervall zwischen den Analysedatensätzen mit Testadministration und –durchführung von Testleiter/-innen bzw. Lehrkräften zu Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik für Prognosen ab der 3. Klasse.

ESM-11.

Vergleichende Darstellung der adjustierten standardisierten Regressionskoeffi zienten (β adj) mit 95 %-Konfidenzintervall zwischen den Analysedatensätzen mit Testadministration und -durchführung von Testleiter/-innen bzw. Lehrkräften zu Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik für Prognosen ab der 3. Klasse.

ESM-12.

Interkorrelationsmatrix aller Analysevariablen für KEGSTeildatensatz (3. Klasse, Testdurchführung von Lehrkräften)

Tabelle ESM-1: Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Leistungstests

Kompetenztest (Prädiktor)	Prognosedauer (Jahre)	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Mathematiktest und Mathematiktest							
Thüringer Kompetenztest (VERA 3)	5	3.- 8.	Thüringer Kompetenztest (VERA 8)	r = .63			Nachtigall, 2014[a]
Thüringer Kompetenztest (VERA 3)	3	3.- 6.	Thüringer Kompetenztest (VERA 6)	r = .67			Nachtigall, 2014[a]
DEMAT 1+	3	1.- 4.	DEMAT 4	r = .59			Hasselhorn, Roick & Gölitz, 2005
DEMAT 1+, 2+	2	1.- 3., 2.- 4.	DEMAT 3+, 4	r = .55, .64			Hasselhorn, Roick & Gölitz, 2005
DEMAT 1+, 2+	1	1.- 2., 2.- 3.	DEMAT 2+, 3+	r = .67, .65			Hasselhorn, Roick & Gölitz, 2005
DEMAT 1+	1	1.- 2.	DEMAT 2+	r = .80			Dornheim, 2007
DEMAT 3+	1	3.- 4.	DEMAT 4	r = .68	Intelligenz 4. Klasse Mathematiknote 3. Klasse	r_{adj} = .47 r_{adj} = .49	Roick, Gölitz & Hasselhorn, 2004
MBK-1	2	1.- 2.	HRT 1-4	r = .71/.64			Sinner, Ennemoser & Krajewski, 2011
MBK-1	1	1.- 2.	DEMAT 1+	r = .72			Sinner, Ennemoser & Krajewski, 2011

(Tabelle ESM-1 wird fortgesetzt)

Tabelle ESM-1 (Fortsetzung): Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Leistungstests

Kompetenztest (Prädiktor)	Prognosedauer (Jahre)	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Deutschtest und Deutschtest							
Thüringer Kompetenztest (VERA 3)	5	3.- 8.	Thüringer Kompetenztest (VERA 8)	r = .64			Nachtigall, 2014[a]
Thüringer Kompetenztest (VERA 3)	3	3.- 6.	Thüringer Kompetenztest (VERA 6)	r = .65			Nachtigall, 2014[a]
HASE	≈ 3	0.- 3.	Knuspel-L, WLLP	r = .51, r = .31			Roos, Schöler & Treutlein, 2007
ELFE 1-6	1	3.- 4.	ELFE 1-6	r = .65[b]			Robitzsch, Dörfler, Pfost & Artelt, 2011
ELFE 1-6	0.5	3.- 4., 4.	ELFE 1-6	r = .75[b],.72[b]			Robitzsch, Dörfler, Pfost & Artelt, 2011

(Tabelle ESM-1 wird fortgesetzt)

Tabelle ESM-1 (Fortsetzung): Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Leistungstests

Kompetenztest (Prädiktor)	Prognosedauer (Jahre)	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Mathematiktest und Mathematiknote							
VERA (Landau)	0	4.		r = -.65[b]			Hochweber, 2010
Thüringer Kompetenztest (VERA 3)	0	3.		r = -.64			Nachtigall, 2014[a]
Thüringer Kompetenztest (VERA 6)	0	6.		r = -.57			Nachtigall, 2014[a]
IQB-Aufgabenpool (BISTA)	0	3.		r = -.48 bis -.67[b,c]			Winkelmann & Robitzsch, 2009
IQB-Aufgabenpool (BISTA)	0	4.		r = -.62 bis -.72[b,c]			Winkelmann & Robitzsch, 2009
DEMAT 2+	2	2.- 4.		r = -.64			Krajewski, Liehm & Schneider, 2004
DEMAT 2+	1	2.- 3.		r = -.67			Krajewski, Liehm & Schneider, 2004
DEMAT 3+	1	3.- 4.		r = -.69	Mathematik-note 3. Klasse	r_{adj} = -.46	Roick, Gölitz & Hasselhorn, 2004

(Tabelle ESM-1 wird fortgesetzt)

Tabelle ESM-1 (Fortsetzung): Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Leistungstests

	Prognosedauer (Jahre)	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Deutschtest und Deutschnote							
IQB-Aufgabenpool (BISTA) - Lesen, Hören, Sprachgebrauch, Rechtschreibung[d]	2	4.- 6.		r = -.38	Intelligenz, HISEI, Migr., Geschlecht, Mathematik (Test, Note), Deutschnote, 4. Klasse	β_{adj} = .18 (Hauptschule), .19 (Realschule), .17 (Gymnasium)	Hildebrandt & Watermann, 2015
Thüringer Kompetenztest (VERA 3) - Deutsch	0	3.		r = -.63			Nachtigall, 2014[a]
Thüringer Kompetenztest (VERA 8) - Deutsch	0	6.		r = -.53			Nachtigall, 2014[a]
IQB-Aufgabenpool (BISTA) - Lesen	0	3.		r = -.62[b]			Bremerich-Vos, Böhme & Robitzsch, 2009
HASE	≈ 1	0.- 1.	Lesen Rechtschreibung	r = -.45 r = -.40			Roos, Schöler & Treutlein, 2007
HASE	≈ 2	0.- 2.	Deutsch Lesen Rechtschreibung	r = -.45 r = -.42 r = -.38			Roos, Schöler & Treutlein, 2007
HASE	≈ 3	0.- 3.	Deutsch Lesen Rechtschreibung	r = -.45 r = -.42 r = -.40			Roos, Schöler & Treutlein, 2007

(Tabelle ESM-1 wird fortgesetzt)

Tabelle ESM-1 (Fortsetzung): Forschungsstand zu bildungsstandardbasierten und kommerziell erhältlichen, standardisierten Leistungstests

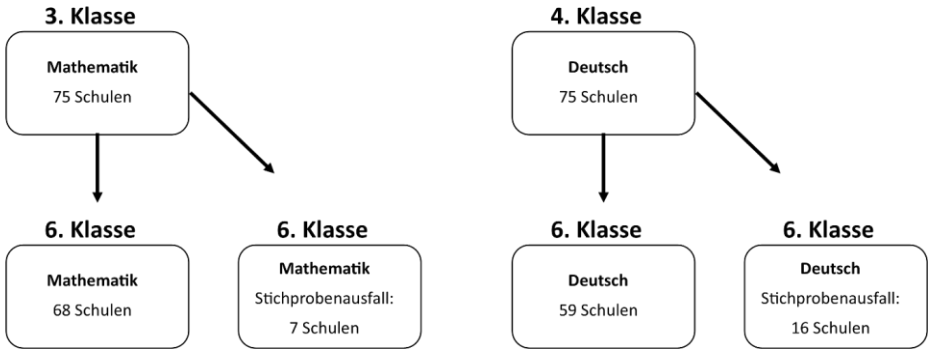
Kompetenztest (Prädiktor)	Prognosedauer (Jahre)	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Mathematiktest und Übertrittsempfehlung							
BISTA-Heft 1 und 2[e]	0	5.	Gymnasial-besuch	r = .53, .55	Intelligenz, Deutsch-Lesetest, Orthografie-test	b = 0.003	Köller, Eßel-Ullmann & Paasch, 2012
DEMAT 3+	1	3.- 4.	Übertritts-empfehlung (Hauptschule, Realschule, Gymnasium)	ρ = .60			Roick, Gölitz & Hasselhorn, 2004
DEMAT 4	0	4.	Übertritts-empfehlung (Hauptschule, Realschule, Gymnasium)	ρ = .67	Mathematik- und Deutschnote	ρ = .20	Gölitz, Roick & Hasselhorn, 2006

Anmerkungen. VERA = Vergleichsarbeiten – für dritte, sechste und achte Klassen (bildungsstandardbasierte Tests). DEMAT = Deutscher Mathematiktest – für erste, zweite, dritte und vierte Klassen. MBK-1 = Test zur Erfassung mathematischer Basiskompetenzen. HRT 1-4 = Heidelberger Rechentest. HASE = Heidelberger Auditives Screening in der Einschulungsuntersuchung. Knuspel-L = Knuspels Leseaufgaben. ELFE 1-6 = Ein Leseverständnistest für Erst- bis Sechstklässler. WLLP = Würzburger Leise-Leseprobe. IQB = Institut zur Qualitätsentwicklung im Bildungswesen. BISTA = Bildungsstandardbasierte Tests bzw. Aufgaben. HISEI = Höchster ISEI-Wert der Eltern. Migr. = Migrationshintergrund. r = Korrelationskoeffizient. r_{adj} = adjustierter Korrelationskoeffizient. β_{adj} = adjustierter Regressionskoeffizient. b = Probit-Koeffizient. ρ = Rangkorrelationskoeffizient nach Spearman.

[a] persönliche Mitteilung der Werte. [b] Latente Modellierung. [c] Separat für Leitideen ermittelt und Zentrierung am Klassenmittelwert für Testleistung und Schulnote. [d] Granzer, Köller & Bremerich-Vos, 2009. [e] Bildungsstandards. Kompetenzen überprüfen. Grundschule Klasse 3/4 - Heft 1, Heft 2 (Granzer et al., 2008a, 2008b).

ESM-2:

Überblick zum Stichprobenausfall der Schulen bis zur 6. Klasse für Mathematik (ab der 3. Klasse) und für Deutsch (ab der 4. Klasse)



	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
M-TL 3	438	50	427	54	-.22	439	50	433	52	-.12
M-TL 4	527	38	508	44	-.49	526	38	517	40	-.23
M-TL 5	568	46	564	90	-.08	570	47	557	57	-.26
D-TL 4	562	40	553	50	-.22	561	42	549	38	-.29
D-TL 5	561	35	536	60	-.66	562	35	549	40	-.37
M-N 4	2.27	0.33	2.41	0.56	.34	2.25	0.33	2.44	0.43	.54
D-N 4	2.10	0.27	2.29	0.32	.69	2.10	0.27	2.19	0.31	.32
M-N 5	2.56	0.39	2.34	0.50	-.55	2.53	0.39	2.60	0.41	.18
D-N 5	2.37	0.36	2.55	0.62	.46	2.35	0.36	2.51	0.44	.42
N2 4	16.93	2.08	16.44	2.04	-.24	17.02	2.11	16.28	1.86	-.36
Eink 3	4.92	0.96	5.11	1.37	.19	4.94	0.93	4.84	1.26	-.10
Buch 3	4.81	0.49	4.94	0.72	.25	4.80	0.47	4.91	0.66	.21
Schul 3	6.58	0.49	6.44	0.76	-.27	6.58	0.48	6.47	0.72	-.20
Beruf 3	3.40	0.69	3.28	1.19	-.16	3.44	0.68	3.18	1.00	-.34
Kultur 3	1.75	0.14	1.75	0.25	.00	1.76	0.14	1.72	0.21	-.26
Güter 3	6.38	0.57	6.47	0.69	.16	6.38	0.56	6.36	0.69	-.03
Mädchen	0.49	0.10	0.52	0.06	.31	0.50	0.10	0.50	0.11	.00

Anmerkungen. In der 1. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung angegeben (bspw. M-TL 3 = Erhebung in der 3. Klasse). M-TL = Mathematiktestleistung; D-TL = Deutschtestleistung. M-N = Mathematiknote; D-N = Deutschnote; N2 = Figurenanalogien; Eink = monatliches Haushaltsnettoeinkommen; Buch = Buchbesitz; Schul = schulischer Abschluss der Eltern; Beruf = beruflicher Abschluss der Eltern; Kultur = kulturelle Aktivitäten; Güter = Summe der Wohlstandsgüter; Mädchen = Anteil der Mädchen. *M* = Mittelwert, *SD* = Standardabweichung, *d* = Cohen's *d* mit gepoolter Standardabweichung. Für die KEGS-Stichprobe konnte in einer Arbeit von Kuhl (2009; s. a. Fuchs & Brunner, 2014, Abb. 5) zum Zeitpunkt der dritten Klasse gezeigt werden, dass es sich um eine repräsentative Stichprobe für Brandenburger Grundschüler/-innen handelt.

ESM-3:

Darstellung der Items zur Erfassung des familiären Hintergrundes[2]

Höchster schulischer Abschluss der Eltern.

Welchen höchsten Schulabschluss haben Sie und gegebenenfalls Ihr mit Ihnen im Haushalt lebende(r) Partner(in)?

Ich …

(1) …habe keine Schule besucht

(2) …bin ohne Abschluss von der Schule abgegangen

(3) …habe einen Abschluss einer Sonder-/Förderschule

(4) …habe einen Abschluss einer Polytechnischen Oberschule nach der 8. Klasse

(5) …habe einen Hauptschulabschluss/Volksschulabschluss

(6) …habe einen Realschulabschluss (mittlere Reife) oder Abschluss der Polytechnischen Oberschule nach der 10. Klasse

(7) …habe eine Fachhochschulreife

(8) …habe ein Abitur/Hochschulreife

(9) …habe einen anderen Schulabschluss (z. B. aus dem Ausland)

Kodierung: 1-9

Anmerkungen: Antwortalternative (9) nicht in Analysen berücksichtigt, $n = 9$

[2] Von der Korrektur grammatikalischer Fehler und Rechtschreibfehler wurde bewusst zugunsten einer transparenten Dokumentation der Erhebungsinstrumente Abstand genommen (S. 7-12).

Höchster beruflicher Abschluss der Eltern.

Welchen beruflichen Ausbildungsabschluss haben Sie?

Ich habe…

(1) …keine abgeschlossene Berufsausbildung

(2) …eine abgeschlossene Lehre/Abschluss an einer Berufsaufbauschule

(3) …einen Abschluss an einer Berufsfachschule oder Handelsschule

(4) …einen Abschluss an einer Fachschule, Meister- oder Techniker-Schule oder Schule des Gesundheitswesens

(5) …einen Fachhochschulabschluss/Diplom (FH), Abschluss einer Berufsakademie

(6) …einen Hochschulabschluss (Magister, Diplom oder Staatsexamen)

(7) …eine Promotion

(8) …einen anderen beruflichen Abschluss (z. B. aus dem Ausland)

Kodierung: 1-8

Anmerkungen: Antwortalternative (8) nicht in Analysen berücksichtigt, $n = 8$

Monatliches Nettoeinkommen des Haushaltes.

Wenn Sie alle Gelder zusammenzählen, die Ihrem Haushalt monatlich zur Verfügung stehen: In welcher Spanne bewegt sich das monatliche Nettoeinkommen Ihres Haushalts?

(1) Unter 500 €

(2) 500 bis 1000 €

(3) 1000 bis 1500 €

(4) 1500bis 2000 €

(5) 2000 bis 2500 €

(6) 2500 bis 3000 €

(7) 3000 bis 4000 €

(8) 4000 bis 6000 €

(9) Über 6000 €

Kodierung: 1-9

Anzahl der Bücher im eigenen Haushalt.

Wie viele Bücher gibt es bei Ihnen zu Hause ungefähr?

(Kreuzen Sie bitte nur ein Kästchen an!)

(1) Keine

(2) 1-10 Bücher

(3) 11-50 Bücher

(4) 51-100 Bücher

(5) 101-250 Bücher

(6) 251-500 Bücher

(7) mehr als 500 Bücher

Zur Erleichterung der Fragenbeantwortung:
Dieses eine Regal enthält etwa 100 Bücher

Kodierung: 1-7

Vorhandensein von Wohlstandsgütern.

Gibt es diese Dinge bei Ihnen zu Hause?

(Kreuzen Sie bitte in jeder Zeile nur ein Kästchen an!)

Gibt es bei Ihnen zu Hause…

(1) ein eigenes Kinderzimmer für das Kind, das an MathePlus teilnimmt?

(2) einen Computer, den Ihr Kind zum Arbeiten für die Schule benutzen kann?

(3) einen Internet-Anschluss?

(4) ein Musikinstrument?

(5) einen Fernseher?

(6) ein Lexikon?

(7) einen eigenen Garten?

(8) eine abonnierte Tageszeitung?

Kodierung: 0 = *nicht angekreuzt*; 1 = *angekreuzt*

Gemeinsame kulturelle Aktivitäten.

Wie häufig haben Sie gemeinsam mit Ihrem Kind im letzten Jahr …

(Kreuzen Sie bitte in jeder Zeile nur ein Kästchen an!)

(1) im Kino einen Film gesehen?

(2) ein Museum oder eine Kunstausstellung besucht?

(3) ein Rock-, Pop- oder ähnliches Konzert besucht?

(4) eine Oper, ein Ballett oder ein klassisches Konzert besucht?

(5) ein Theater besucht?

(6) eine Sportveranstaltung besucht (z. B. ein Fußballspiel im Stadion)?

Kodierung: 1 = *Nie*; 2 = *Etwa 1- bis 2-mal*; 3 = *3- bis 4-mal*; 4 = *Mehr als 4-mal*

ESM-4:

„Verlinkung“ der bildungsstandardbasierten Kompetenztests

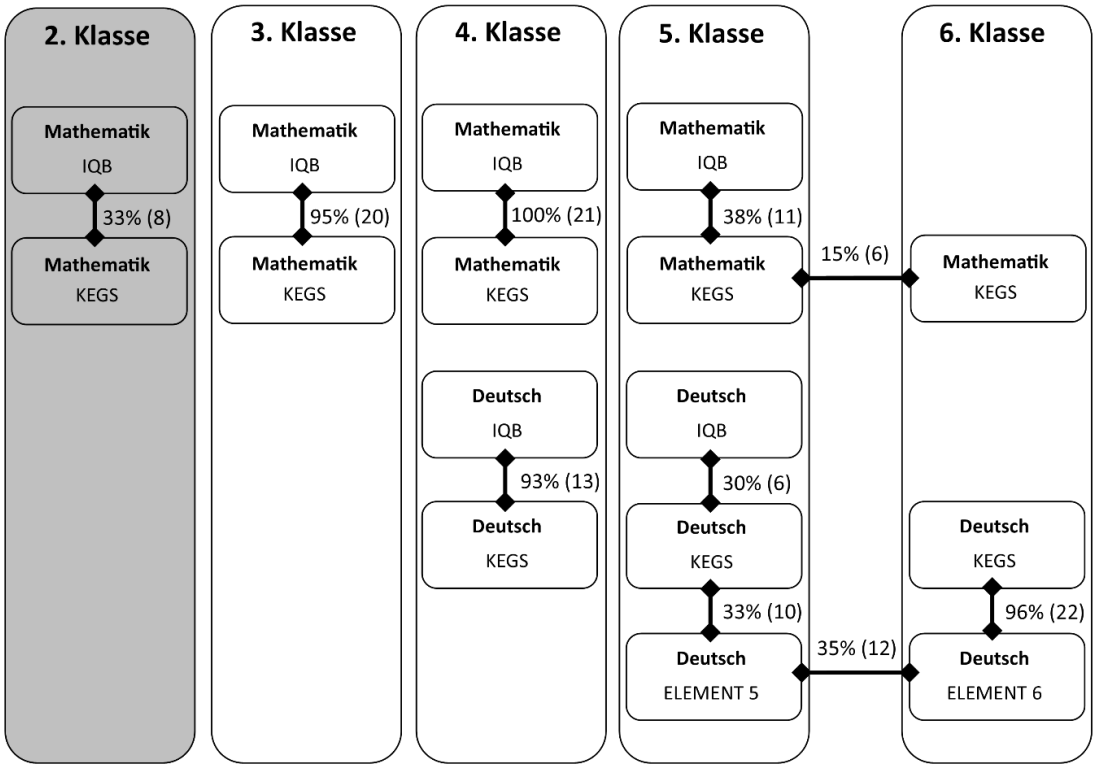


Abbildung ESM-4. Veranschaulichung der Messinvarianz „Verlinkung“ der bildungsstandardbasierten Kompetenztests in Mathematik und Deutsch (Lesen) mit Itemparametern aus dem IQB-Aufgabenpool und der ELEMENT-Studie in Klasse 5 und 6. Die Daten in der 2. Klasse wurden nicht in die Analysemodelle dieses Artikels einbezogen (grau hinterlegt). Die Prozentangaben beziehen sich auf den Anteil der „verlinkten“ Aufgaben am fachspezifischen KEGS-Test, für den Itemparameter aus den Referenzstudien vorlagen. In Klammern steht die Anzahl an sogenannten „Linking“-Items. Lesebeispiel der Abbildung: Für 95 % der Items (also 20 Items) für den Mathematiktest in der dritten Klasse lagen Itemparameter aus der Normierungsstudie des IQB vor.

ESM-5: Interkorrelationsmatrix aller Analysevariablen

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]
[1]	M-TL 3	–	.56	.63	.59	.37	.36	.40	-.54	-.44	-.48	-.40	-.51	-.45	.32	.18	.21	.29	.24	.25	*.11*	.23	.53
[2]	M-TL 4		–	.66	.62	.49	.50	.55	-.59	-.55	-.57	-.47	-.58	-.55	.38	.19	.26	.33	.26	.28	.21	.34	.68
[3]	M-TL 5			–	.75	.43	.47	.50	-.63	-.58	-.64	-.54	-.67	-.57	.34	.35	.28	.30	.32	.29	*.15*	.26	.72
[4]	M-TL 6				–	.44	.45	.53	-.61	-.53	-.60	-.51	-.65	-.58	.37	.38	.24	.32	.30	.32	.19	.27	.72
[5]	D-TL 4					–	.60	.60	-.43	-.49	-.45	-.47	-.45	-.50	.30	.26	.13	.29	.20	.25	*.12*	.25	.62
[6]	D-TL 5						–	.61	-.40	-.49	-.44	-.50	-.45	-.56	.30	.30	.25	.30	.28	.32	.15	.29	.64
[7]	D-TL 6							–	-.41	-.48	-.49	-.47	-.43	-.57	.32	.28	.17	.36	.36	.29	.14	.38	.65
[8]	M-N 4								–	.58	.71	.54	.70	.58	-.38	-.29	-.31	-.30	-.25	-.27	-.17	-.31	-.70
[9]	D- N 4									–	.63	.71	.60	.67	-.27	-.37	-.22	-.29	-.31	-.31	*-.13*	-.29	-.78
[10]	M-N 5										–	.61	.72	.63	-.34	-.32	-.27	-.30	-.30	-.25	*-.13*	-.31	-.76
[11]	D-N 5											–	.57	.71	-.24	-.39	-.28	-.26	-.22	-.22	*-.02*	-.29	-.80
[12]	M-N 6												–	.69	-.41	-.32	-.28	-.34	-.25	-.28	*-.12*	-.23	-.89
[13]	D-N 6													–	-.34	-.42	-.34	-.35	-.34	-.32	*-.12*	-.27	-.90
[14]	N2 4														–	.21	*.05*	.23	.24	.18	*.07*	.15	.39
[15]	HISEI 6															–	.44	.38	.34	.52	*.11*	.23	.44
[16]	Eink 3																–	.46	.49	.52	.32	.52	.35
[17]	Buch 3																	–	.63	.62	.35	.48	.40
[18]	Schul 3																		–	.68	.34	.43	.32
[19]	Beruf 3																			–	.38	.44	.31
[20]	Kultur 3																				–	.36	*.09*
[21]	Güter 3																					–	.37
[22]	Gym 6																						–
	M	420	521	566	589	553	562	631	2.29	2.09	2.50	2.41	2.46	2.35	16.77	49.40	5.38	4.83	6.73	3.88	1.78	6.27	0.56
	SD	111	108	118	123	121	105	120	0.92	0.83	1.01	0.84	1.05	0.89	6.75	16.02	2.05	1.40	1.38	1.70	0.48	1.46	–
	Mis (%)	19	17	22	24	15	22	29	20	20	22	22	24	24	16	45	54	48	51	51	56	57	25

Anmerkungen. In der 1. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung angegeben (bspw. M-TL 3 = Erhebung in der 3. Klasse). M-TL = Mathematiktestleistung; D-TL = Deutsch-testleistung. M-N = Mathematiknote; D-N = Deutschnote; N2 = Figurenanalogien; HISEI = Höchster ISEI-Wert der Eltern; Eink = monatliches Haushaltsnettoeinkommen; Buch = Buchbesitz; Schul = schulischer Abschluss der Eltern; Beruf = beruflicher Abschluss der Eltern; Kultur = kulturelle Aktivitäten; Güter = Summe der Wohlstandsgüter; Gym = Gymnasialempfehlung. Alle Korrelationen (außer kursiv gesetzte Werte) sind signifikant $p < .05$ (zweiseitig). *M* = Mittelwert, *SD* = Standardabweichung, Mis (%) = Anteil fehlender Werte in Prozent.

ESM-6: Replikation der Analysen – Schulnoten klassenzentriert

ESM-6.1-Tabelle: *Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik – Schulnoten klassenzentriert*

		Vorhersage mit Mathematik-Kompetenztest in der 3. Klasse auf Kriterien in der						Vorhersage mit Mathematik-Kompetenztest in der 4. Klasse auf Kriterien in der			
Modell		4. Klasse		5. Klasse		6. Klasse		5. Klasse		6. Klasse	
Kriterium: Testleistung in Mathematik											
(a)	β	.56	[.49, .64]	.63	[.57, .69]	.58	[.50, .66]	.66	[.59, .72]	.61	[.55, .70]
	*B**70	38	[33, 44]	47	[42, 52]	45	[38, 52]	50	[43, 58]	48	[41, 56]
	R^2	.32		.40		.34		.43		.38	
(b)	β_{adj}	.32[a]	[.22, .41][a]	.42[a]	[.34, .51][a]	.34[a]	[.26, .49][a]	.47	[.37, .57]	.38	[.28, .48]
	B_{adj}*70	22[a]	[15, 28][a]	31[a]	[25, 38][a]	26[a]	[18, 34][a]	36	[28, 44]	30	[22, 38]
	R^2	.47[a]		.55[a]		.51[a]		.56		.52	
	sr^2	.06[a]		.12[a]		.08[a]		.13		.09	
Kriterium: Mathematiknote											
(a)	β	-.53	[-.62, -.43]	-.46	[-.55, -.36]	-.50	[-.59, -.40]	-.53	[-.62, -.45]	-.54	[-.62, -.46]
	*B**70	-.28	[-.33, -.23]	-.28	[-.35, -.21]	-.28	[-.33, -.23]	-.35	[-.42, -.28]	-.35	[-.40, -.30]
	R^2	.28		.21		.25		.28		.30	
(b)	β_{adj}	-.41[b]	[-.51, -.31][b]	-.07[a]	[-.14, .01] [a]	-.13[a]	[-.22, -.04][a]	-.16	[-.25 ,-.07]	-.19	[-.27, -.11]
	B_{adj}*70	-.21[b]	[-.26, -.16][b]	-.07[a]	[-.15, .01] [a]	-.07[a]	[-.12, -.02][a]	-.07	[-.11, -.03]	-.14	[-.19, .09]
	R^2	.38[b]		.60[a]		.57[a]		.61		.58	
	sr^2	.14[b]		.01[a]		.01[a]		.02		.02	

Anmerkungen. Modell a: bivariates Regressionsmodell. Modell b: multiples Regressionsmodell (Spezifikation s. Text). In Klammern wird das 95%-Konfidenzintervall ausgewiesen. *B**70, B_{adj}*70 Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. R^2 = Determinationskoeffizient. sr^2 = Quadrierter Semipartialkorrelationskoeffizient.

[a] In der 3. Klasse wurden die Schulnoten nicht erfasst; in Modell b wurde daher für die Mathematiknote der 4. Klasse kontrolliert. [b] In der 3. Klasse wurden die Schulnoten nicht erfasst; in Modell b wurde daher nicht für die Mathematiknote kontrolliert.

ESM-6.2-Tabelle: *Forschungsfrage 2: Vorhersage zukünftiger Testleistungen und Schulnoten in Deutsch – Schulnoten klassenzentriert*

		Vorhersage mit Deutsch-Lesekompetenztest in der 4. Klasse auf Kriterien in der			
Modell		5. Klasse		6. Klasse	
Kriterium: Testleistung in Deutsch-Lesen					
(a)	β	.60	[.53, .67]	.61	[.54, .68]
	B*70	36	[31, 42]	42	[37, 48]
	R^2	.36		.37	
(b)	β_{adj}	.45	[.36, .55]	.44	[.35, .53]
	B_{adj}*70	28	[21, 34]	31	[24, 37]
	R^2	.44		.51	
	sr^2	.15		.13	
Kriterium: Deutschnote					
(a)	β	-.45	[-.54, -.36]	-.49	[-.56, -.41]
	B*70	-.21	[-.25, -.17]	-.21	[-.25, -.17]
	R^2	.20		.24	
(b)	β_{adj}	-.13	[-.20, -.05]	-.19	[-.27, -.11]
	B_{adj} *70	-.07	[-.11, -.03]	-.07	[-.10, -.04]
	R^2	.61		.57	
	sr^2	.01		.01	

Anmerkungen. Modell a: bivariates Regressionsmodell. Modell b: multiples Regressionsmodell (Spezifikation s. Text). In Klammern wird das 95%-Konfidenzintervall ausgewiesen. B*70, B_{adj}*70 Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. R^2 = Determinationskoeffizient. sr^2 = Quadrierter Semipartialkorrelationskoeffizient.

ESM-6.3-Tabelle: *Forschungsfrage 3: Vorhersage der Gymnasialempfehlung in der 6. Klasse auf Grundlage bildungsstandardbasierter Testleistungen in Mathematik und Deutsch in der 4. Klasse – Schulnoten klassenzentriert*

Modell	Testdomäne	Koeffizient	95%-Konfidenzintervall
(a)	Mathematik		
	$Exp(\beta)_{adj}$	3.48	[2.45, 4.96]
	$Exp(B*70)_{adj}$	2.32	[1.76, 3.05]
	Deutsch		
	$Exp(\beta)_{adj}$	2.47	[1.80, 3.40]
	$Exp(B*70)_{adj}$	1.75	[1.53, 2.01]
(b)	Mathematik		
	$Exp(\beta)_{adj}$	2.08	[1.27, 3.42]
	$Exp(B*70)_{adj}$	1.63	[1.24, 2.15]
	Deutsch		
	$Exp(\beta)_{adj}$	1.83	[1.24, 2.71]
	$Exp(B*70)_{adj}$	1.42	[1.08, 1.87]

Anmerkungen. Modell a: multiples Regressionsmodell mit den Testleistungen in Deutsch und Mathematik als Prädiktoren (Spezifikation s. Text). Modell b: multiples Regressionsmodell, in dem zusätzlich zu den Testleistungen in Deutsch und Mathematik für zahlreiche leistungsrelevante Kovariaten kontrolliert wurde (Spezifikation s. Text). Exp(β) „Semistandardisierter" Wert, Vorhersagekraft für einen Leistungsunterschied von einer Standardabweichung des Prädiktors, Exp(*B**70) Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. Die Werte in der Spalte „Koeffizient" können auch als Odds Ratios interpretiert werden.

ESM-7:

Replikation der Analysen mit KEGS-Teildatensatz (3. Klasse, Testadministration und -durchführung durch Lehrkräfte)

ESM-7.1-Tabelle: *Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik – Testdurchführung von Lehrkräften (3.Klasse)*

		Vorhersage mit Mathematik-Kompetenztest in der 3. Klasse auf Kriterien in der					
Modell		4. Klasse		5. Klasse		6. Klasse	
Kriterium: Testleistung in Mathematik							
(a)	β	.61	[.51, .71]	.62	[.50, .75]	.65	[.57, .74]
	B*70	40	[33, 47]	49	[38, 60]	49	[40, 57]
	R^2	.38		.39		.43	
(b)	β_{adj}	.37[a]	[.22, .52][a]	.39[a]	[.24, .55][a]	.44[a]	[.30, .62][a]
	B_{adj}*70	24[a]	[15, 34][a]	31[a]	[19, 43][a]	32[a]	[21, 44][a]
	R^2	.51[a]		.53[a]		.54[a]	
	sr^2	.09[a]		.10[a]		.12[a]	
Kriterium: Mathematiknote							
(a)	β	-.52	[-.65, -.39]	-.58	[-.68, -.47]	-.59	[-.68, -.50]
	B*70	-.28	[-.42, -.14]	-.35	[-.49, -.21]	-.35	[-.49, -.21]
	R^2	.27		.33		.35	
(b)	β_{adj}	-.48[b]	[-.60, -.36][b]	-.26[a]	[-.40, -.13] [a]	-.32[a]	[-.44, -.21][a]
	B_{adj}*70	-.28[b]	[-.42, -.14][b]	-.14[a]	[-.28, .00] [a]	-.21[a]	[-.35, -.07][a]
	R^2	.32[b]		.57[a]		.51[a]	
	sr^2	.18[b]		.04[a]		.07[a]	

Anmerkungen. Modell a: bivariates Regressionsmodell. Modell b: multiples Regressionsmodell (Spezifikation s. Text). In Klammern wird das 95%-Konfidenzintervall ausgewiesen. B*70, B_{adj}*70 Vorhersagekraft für einen Leistungsunterschied von einer Kompetenzstufe. R^2 = Determinationskoeffizient. sr^2 = Quadrierter Semipartialkorrelationskoeffizient.

[a] In der 3. Klasse wurden die Schulnoten nicht erfasst; in Modell b wurde daher für die Mathematiknote der 4. Klasse kontrolliert. [b] In der 3. Klasse wurden die Schulnoten nicht erfasst; in Modell b wurde daher nicht für die Mathematiknote kontrolliert.

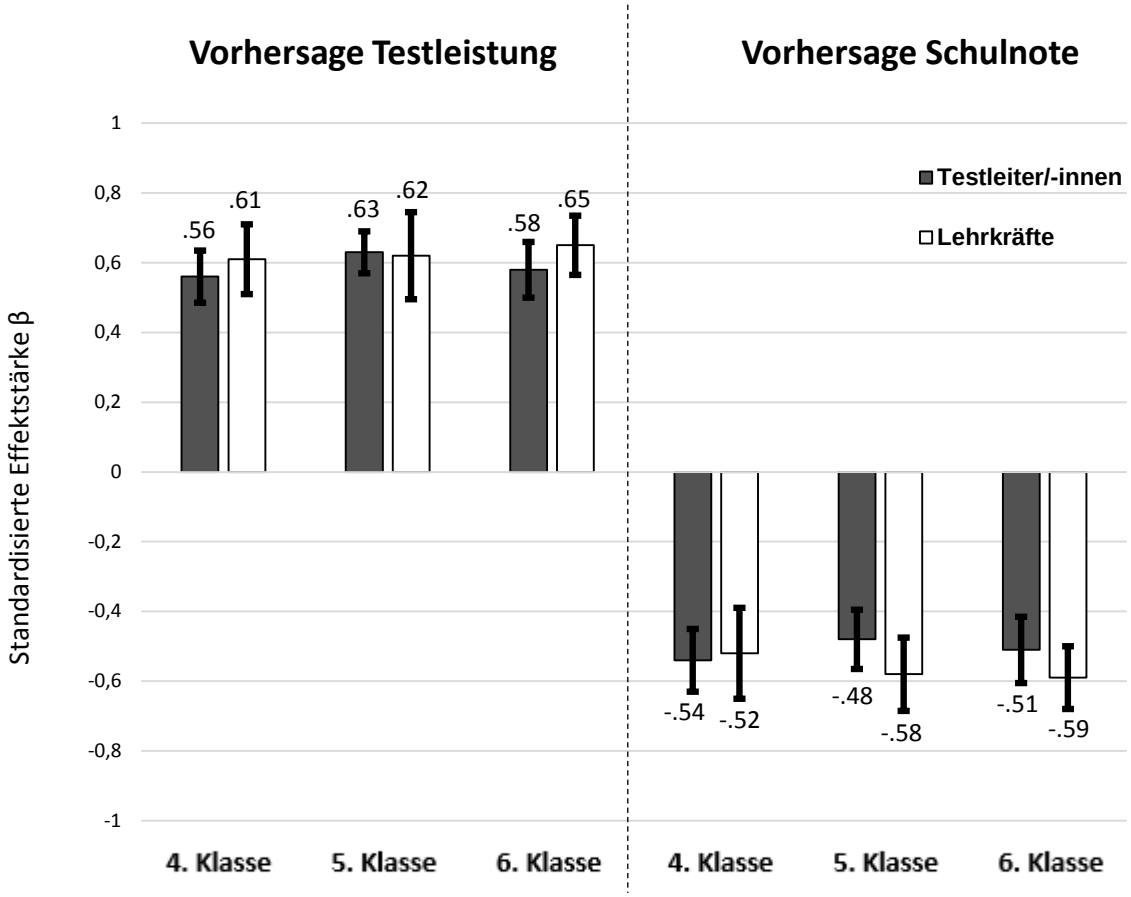


Abbildung ESM-7.2. Vergleich der standardisierten Regressionskoeffizienten (β) mit 95 %-Konfidenzintervall zu Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik für Prognosen ab der 3. Klasse. Testleiter/-innen: Werte für den Analysedatensatz mit Testadministration und -durchführung zu allen Erhebungszeitpunkten durch Testleiter/-innen. Lehrkräfte: Werte für KEGS-Teildatensatz mit Testadministration und -durchführung in der 3. Klasse durch Lehrkräfte.

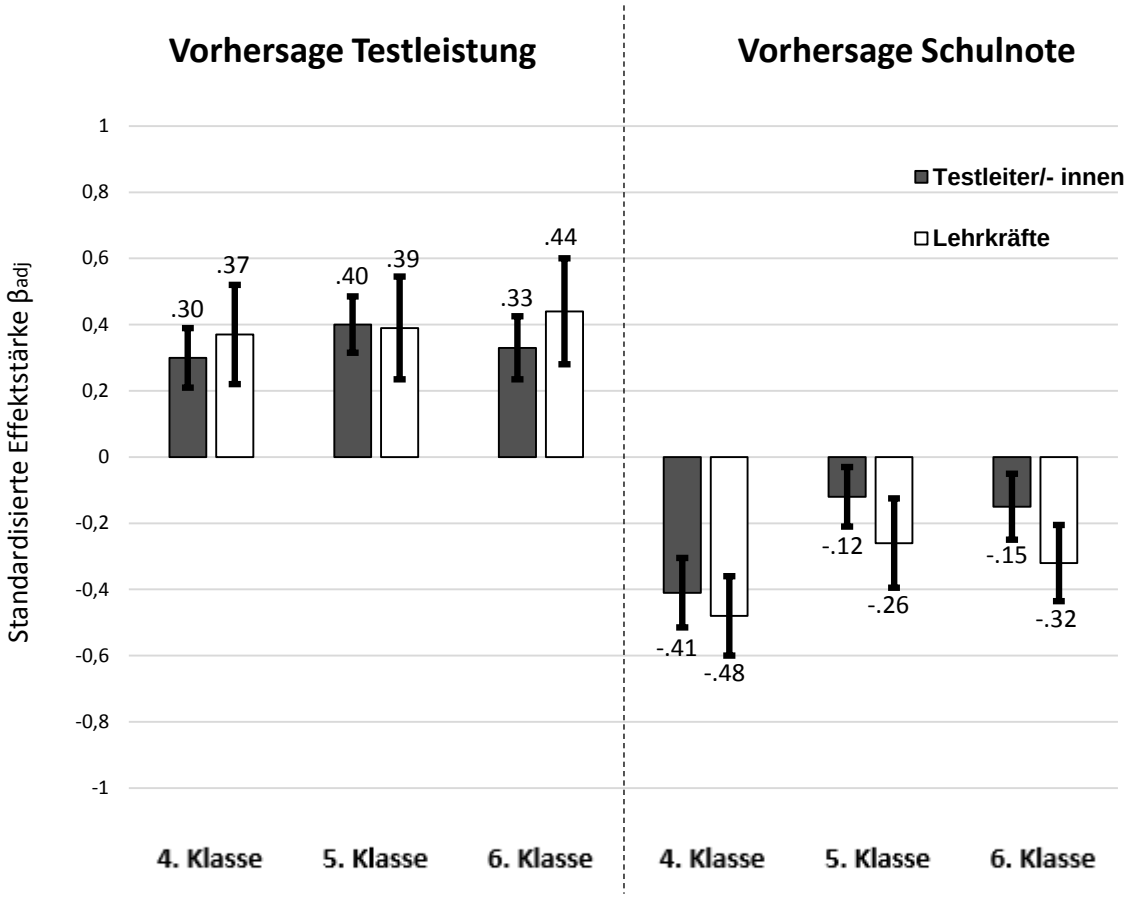


Abbildung ESM-7.3. Vergleichende Darstellung der adjustierten standardisierten Regressionskoeffizienten (β_{adj}) mit 95 %-Konfidenzintervall zur Forschungsfrage 1: Vorhersage zukünftiger Testleistungen und Schulnoten in Mathematik für Prognosen ab der 3. Klasse. Testleiter/-innen: Werte für den Analysedatensatz mit Testadministration und -durchführung zu allen Erhebungszeitpunkten durch Testleiter/-innen. Lehrkräfte: Werte für KEGS-Teildatensatz mit Testadministration und -durchführung in der 3. Klasse durch Lehrkräfte.

ESM-7.4: Interkorrelationsmatrix aller Analysevariablen für KEGS-Teildatensatz (3. Klasse, Testdurchführung von Lehrkräften)

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]
[1]	M-TL 3	–	.56	.63	.59	.37	.36	.40	-.54	-.44	-.48	-.40	-.51	-.45	.32	.18	.21	.29	.24	.25	*.11*	.23	.53
[2]	M-TL 4	.61	–	.66	.62	.49	.50	.55	-.59	-.55	-.57	-.47	-.58	-.55	.38	.19	.26	.33	.26	.28	.21	.34	.68
[3]	M-TL 5	.62	.57	–	.75	.43	.47	.50	-.63	-.58	-.64	-.54	-.67	-.57	.34	.35	.28	.30	.32	.29	*.15*	.26	.72
[4]	M-TL 6	.65	.64	.73	–	.44	.45	.53	-.61	-.53	-.60	-.51	-.65	-.58	.37	.38	.24	.32	.30	.32	.19	.27	.72
[5]	D-TL 4	.28	.42	.27	.36	–	.60	.60	-.43	-.49	-.45	-.47	-.45	-.50	.30	.26	.13	.29	.20	.25	*.12*	.25	.62
[6]	D-TL 5	.35	.40	.31	.41	.48	–	.61	-.40	-.49	-.44	-.50	-.45	-.56	.30	.30	.25	.30	.28	.32	.15	.29	.64
[7]	D-TL 6	.41	.44	.43	.51	.53	.53	–	-.41	-.48	-.49	-.47	-.43	-.57	.32	.28	.17	.36	.36	.29	.14	.38	.65
[8]	M-N 4	-.52	-.57	-.53	-.55	-.42	-.42	-.40	–	.58	.71	.54	.70	.58	-.38	-.29	-.31	-.30	-.25	-.27	-.17	-.31	-.70
[9]	D-N 4	-.32	-.46	-.36	-.40	-.50	-.46	-.54	.60	–	.63	.71	.60	.67	-.27	-.37	-.22	-.29	-.31	-.31	*-.13*	-.29	-.78
[10]	M-N 5	-.58	-.60	-.61	-.63	-.45	-.42	-.49	.67	.52	–	.61	.72	.63	-.34	-.32	-.27	-.30	-.30	-.25	*-.13*	-.31	-.76
[11]	D-N 5	-.33	-.39	-.41	-.42	-.50	-.42	-.49	.54	.65	.57	–	.57	.71	-.24	-.39	-.28	-.26	-.22	-.22	*-.02*	-.29	-.80
[12]	M-N 6	-.59	-.62	-.66	-.70	-.35	-.37	-.50	.59	.50	.70	.58	–	.69	-.41	-.32	-.28	-.34	-.25	-.28	*-.12*	-.23	-.89
[13]	D-N 6	-.35	-.41	-.42	-.50	-.48	-.47	-.53	.47	.65	.55	.68	.61	–	-.34	-.42	-.34	-.35	-.34	-.32	*-.12*	-.27	-.90
[14]	N2 4	.34	.35	.29	.29	.23	.22	.24	-.24	-.24	-.34	-.20	-.30	-.26	–	.21	*.05*	.23	.24	.18	*.07*	.15	.39
[15]	HISEI 6	.26	.22	.39	.32	*.07*	.16	.26	-.18	*-.16*	-.26	-.23	-.30	-.27	*.07*	–	.44	.38	.34	.52	*.11*	.23	.44
[16]	Eink 3	.15	*.10*	.21	.18	*.04*	*-.02*	*.05*	*-.11*	*-.08*	*-.05*	*-.14*	-.18	*-.10*	*.07*	.30	–	.46	.49	.52	.32	.52	.35
[17]	Buch 3	.19	*.14*	.17	.16	*.11*	*-.03*	.17	-.17	*-.11*	-.13	*-.11*	-.22	*-.18*	*.05*	.28	.41	–	.63	.62	.35	.48	.40
[18]	Schul 3	*.11*	*.06*	*.11*	*.08*	*.02*	*-.04*	*.13*	*-.10*	*-.07*	-.15	*-.11*	-.18	*-.18*	*.06*	.41	.22	.49	–	.68	.34	.43	.32
[19]	Beruf 3	*.12*	*.10*	*.07*	*.07*	*.08*	*-.00*	*.07*	*-.07*	*-.08*	*-.08*	*-.04*	*-.14*	*-.13*	*-.04*	.37	.37	.53	.62	–	.38	.44	.31
[20]	Kultur 3	*.11*	*.09*	*.12*	*.12*	.20	*.10*	.25	-.19	*-.20*	*-.09*	*-.12*	*-.14*	*-.14*	*.07*	*.15*	*.11*	.31	.25	.33	–	.36	*.09*
[21]	Güter 3	.18	*.13*	.22	.21	*.13*	*.08*	.24	*-.11*	-.25	*-.12*	*-.12*	-.26	*-.16*	*.08*	.22	.49	.48	.36	.44	.35	–	.37
[22]	Gym 6	.48	.63	.63	.70	.50	.48	.60	-.68	-.68	-.72	-.74	-.88	-.84	.27	.41	.20	.25	.28	.22	.23	.23	–
	M	420	521	566	589	553	562	631	2.29	2.09	2.50	2.41	2.46	2.35	16.77	49.40	5.38	4.83	6.73	3.88	1.78	6.27	0.56
	M_{LK}	433	517	568	598	557	575	633	2.23	2.09	2.57	2.49	2.48	2.48	17.37	47.77	4.90	4.86	6.54	3.56	1.78	6.55	0.47
	SD	111	108	118	123	121	105	120	0.92	0.83	1.01	0.84	1.05	0.89	6.75	16.02	2.05	1.40	1.38	1.70	0.48	1.46	–
	SD_{LK}	106	100	120	113	103	94	115	0.88	0.77	0.95	0.88	0.97	0.86	6.41	15.56	1.89	1.44	1.13	1.67	0.52	1.47	–
	Mis (%)	19	17	22	24	15	22	29	20	20	22	22	24	24	16	45	54	48	51	51	56	57	25
	Mis_{LK} (%)	19	13	16	14	11	15	21	17	16	24	24	8	8	11	25	40	32	40	41	49	51	10

Anmerkungen. Korrelationswerte für KEGS-Teildatensatz (3. Klasse Testdurchführung von Lehrkräften) unterhalb der Hauptdiagonalen. Testdurchführung durch geschulte Testleiter/-innen über der Hauptdiagonalen. In der 1. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung angegeben (bspw. M-TL 3 = Erhebung in der 3. Klasse). M-TL = Mathematiktestleistung; D-TL = Deutsch-testleistung. M-N = Mathematiknote; D-N = Deutschnote; N2 = Figurenanalogien; HISEI = Höchster ISEI-Wert der Eltern; Eink = monatliches Haushaltsnettoeinkommen; Buch = Buchbesitz; Schul = schulischer Abschluss der Eltern; Beruf = beruflicher Abschluss der Eltern; Kultur = kulturelle Aktivitäten; Güter = Summe der Wohlstandsgüter; Gym = Gymnasialempfehlung. Alle Korrelationen (außer kursiv gesetzte Werte) sind signifikant $p < .05$ (zweiseitig). *M* = Mittelwert, *SD* = Standardabweichung, Mis (%) = Anteil fehlender Werte in Prozent.

LK Kennwerte für KEGS-Teildatensatz (3. Klasse, Testdurchführung von Lehrkräften).

ESM-1 bis 7: Literatur

Dornheim, D. (2007). *Prädiktion von Rechenleistung und Rechenschwäche. Der Beitrag von Zahlen-Vorwissen und allgemein-kognitiven Fähigkeiten*. Berlin: Logos Verlag Berlin GmbH.

Fuchs, G. & Brunner, M. (2016). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie*, Manuskript in Überarbeitung.

Gölitz, D., Roick, T. & Hasselhorn, M. (2006). *DEMAT 4. Deutscher Mathematiktest für vierte Klassen*. Göttingen: Hogrefe.

Granzer, D., Köller, O., Reiss, K., Robitzsch, A., Walther, G. & Winkelmann, H. (2008a). *Bildungsstandards. Kompetenzen überprüfen. Grundschule. Klasse 3/4 – Heft 1*. Berlin: Cornelsen Verlag.

Granzer, D., Köller, O., Reiss, K., Robitzsch, A., Walther, G. & Winkelmann, H. (2008b). *Bildungsstandards. Kompetenzen überprüfen. Grundschule. Klasse 3/4 – Heft 2*. Berlin: Cornelsen Verlag.

Granzer, D., Köller, O., & Bremerich-Vos, A. (2009). Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule. Weinheim: Belz.

Hasselhorn, M., Roick, T. & Gölitz, D. (2005). Stabilitäten und prognostische Validitäten der Mathematikleistungen. Eine Längsschnittstudie mit der DEMAT-Reihe in der Grundschule. In M. Hasselhorn, H. Marx & W.Schneider (Hrsg.), *Diagnostik von Mathematikleistungen* (Band 4, S. 187–198). Göttingen: Hogrefe.

Hildebrandt, J. & Watermann, R. (2015, März). *Prognostische Validität von curricular gemessenen Testleistungen am Ende der Grundschulzeit*. Vortrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung, Bochum.

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe* (Pädagogische Psychologie und Entwicklungspsychologie) (Band 79). Münster: Waxmann.

Köller, O., Eβel-Ullmann, G. & Paasch, D. (2012). Validierung eines Instruments zur Erfassung Standard-basierter mathematischer Kompetenzen in der Grundschule. = Validation of an instrument on standard-based mathematical competencies in primary school. *Psychologie in Erziehung und Unterricht, 59* (3), 177–190.

Kuhl, P. (2009). *KEGS. Kompetenzentwicklung in der Grundschule in Brandenburg. Zwischenbericht zu den Erhebungen 2007–2009 (Unveröffentlichter Bericht)*. Berlin: Institut für Schulqualität der Länder Berlin und Brandenburg e.V. (ISQ).

Krajewski, K., Liehm, S. & Schneider, W. (2004). *DEMAT 2+. Deutscher Mathematiktest für zweite Klassen*. Göttingen: Beltz.

Nachtigall, C. (2014). *Landesbericht. Thüringer Kompetenztests 2014*. Zugriff am 14.11.2014. Verfügbar unter: https://www.kompetenztest.de/downloads/kompetenztests

Robitzsch, A., Dörfler, T., Pfost, M. & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *43* (4), 213–227.

Roick, T., Gölitz, D. & Hasselhorn, M. (2004). *DEMAT 3+. Deutscher Mathematiktest für dritte Klassen*. Göttingen: Beltz.

Roos, J., Schöler, H. & Treutlein, A. (2007). *Zur prognostischen Validität des Heidelberger Auditiven Screenings in der Einschulungsdiagnostik HASE. Abschlussbericht des Projektes EVER* (S. 33). Heidelberg: Pädagogische Hochschule Heidelberg. Zugriff am 5.8.2015. Verfügbar unter: http://www.ph-heidelberg.de/wp/schoeler/Datein/Abschlussbericht_EVER-HASE_Feb-2007.pdf

Sinner, D., Ennemoser, M. & Krajewski, K. (2011). Entwicklungspsychologische Frühdiagnostik mathematischer Basiskompetenzen im Kindergarten- und frühen Grundschulalter (MBK-0 und MBK-1). In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen* (Band 9, S. 109–126). Göttingen: Hogrefe Verlag GmbH & Co. KG.

Winkelmann, A. & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen- Empirische Befunde zur Dimensionalität. *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 169–196). Weinheim: Beltz Verlag.

Anhang B: Manuskript und elektronische Supplemente - Studie II

STUDIE II: Wie stark variiert die Güte bildungsstandardbasierter Tests zur Prognose zukünftiger schulischer Leistungen zwischen Schulen? – Ein metanalytischer Ansatz zur Validitätsgeneralisierung

Die Teilstudie ist als Zeitschriftenbeitrag eingereicht:

Fuchs, G., Nachtigall, C. & Brunner, M. *(eingereichte Manuskriptfassung)*. Wie stark variiert die Güte bildungsstandardbasierter Tests zur Prognose zukünftiger schulischer Leistungen zwischen Schulen? – Ein metanalytischer Ansatz zur Validitätsgeneralisierung, *Diagnostica*.

(Diese Artikelfassung wurde in dieser Form noch nicht für eine Veröffentlichung angenommen.)

Zusammenfassung

Bildungsstandardbasierte Tests (z. B. die Vergleichsarbeiten [VERA]) können den zukünftigen Schulerfolg vorhersagen. Diese Befunde wurden bislang jedoch nur für Gesamtstichproben ermittelt und es wurde nicht analysiert, inwiefern sich diese Befunde auf Einzelschulen generalisieren lassen. Deshalb haben wir die Validitätsgeneralisierung für VERA-Tests in der 6. Klasse auf VERA-Testergebnisse und Schulnoten im Fach Mathematik in der 8. Klasse für 223 Schulen mit mehreren Bildungsgängen (SMBG; SuS = 6 428) und 78 Gymnasien (GY; SuS = 4 626) metaanalytisch untersucht. Über alle SMBG/GY betrug die mittlere Prognosegüte auf Tests β = .66/.60 und Noten β = .57/.56. Auch bei Kontrolle der Halbjahresnoten in der 6. Klasse leisteten die Tests einen inkrementellen Beitrag zur Prognose. Die Prognosegüte variierte zwischen den Schulen. Diese Heterogenität konnte teilweise durch die schulspezifische Reliabilität der Tests, die Leistungsheterogenität sowie das mittlere Leistungsniveau der Schulen erklärt werden. Das Potential bildungsstandardbasierter Tests zur Leistungsdiagnostik an den Schulen wird diskutiert.

Schlüsselwörter: Leistungsdiagnostik; Schulunterschiede; Prognosegüte; Vergleichsarbeiten; Bildungsstandards

Abstract

Standard-based tests (e.g. VERA) are used in Germany to inform about the future school success of students. The empirical evidence to date is based on analyses for students as a whole ignoring the fact that the students corresponding to different schools. Therefore we examined on the basis of a meta-analyze approach if the prognostic validity of VERA-tests in 6th grade on VERA-tests and grades in math of 8th class can be generalized between 6 428 schools where the students can attend different school tracks (= SMBG; students = 6 428) and 78 schools where they attend the highest school track (= GY, students = 4 626). Analyses showed an averaging prognostic validity between SMBG/GY for later test scores of β = .66/.60 and for later grades of β = .57/.56. The predictive power for all criteria of school success remained even when controlling for grades in the school report (in the 6th grade). There was heterogeneity between schools in prognostic validity. This heterogeneity was to some extend accounted for by the averaging school-level in reliability of the test, the heterogeneity of test performance and for the test performance level. The potential of standard-based tests in state-wide assessment programs for diagnostic purposes of student performance at schools were discussed.

Keywords: diagnostic analyses of mathematics achievement; differences between schools; (prognostic) validity; standard of education

Innerhalb der pädagogisch-psychologischen Diagnostik stellt seit den 1970er-Jahren die kriteriumsorientierte Leistungsmessung ein zentrales Forschungsgebiet dar (Ingenkamp & Lissmann, 2008). Mit den Bildungsstandards (Kultusministerkonferenz [KMK], 2015) liegt nun seit einigen Jahren eine verbindliche, bundesweit einheitliche kriteriale Bezugsnorm vor. Bildungsstandardbasierte Tests, also Tests, bei deren Testentwicklung die Bildungsstandards den theoretischen Rahmen vorgeben, sind weit verbreitet: So sind an allen öffentlichen Schulen in Deutschland Lehrkräfte der 3. und 8. Jahrgangsstufe verpflichtet, jährlich in mindestens einem Fach bildungsstandardbasierte Tests – in Form der Vergleichsarbeiten (VERA) – für ihre insgesamt etwa 1.4 Millionen Schülerinnen und Schüler (SuS) durchzuführen (KMK, 2012 Fassung von 2018; Statistisches Bundesamt, 2017). Konkret formulieren die Bildungsstandards „fachliche und fachübergreifende Basisqualifikationen, die für die weitere schulische und berufliche Ausbildung von Bedeutung sind und die anschlussfähiges Lernen ermöglichen“ (KMK, 2004, S. 7). Dieser Anspruch impliziert, dass auf der Grundlage der Ergebnisse bildungsstandardbasierter Tests, wie bspw. VERA-Tests, zentrale Kriterien des Schulerfolgs hinreichend gut prognostizierbar sein sollten. Bisherige empirische Evidenz zur Prognosegüte (z. B. Fuchs & Brunner, 2017; Graf, Harych, Wendt, Emmrich & Brunner, 2016) weist darauf hin, dass bildungsstandardbasierte Tests bis zu 5 Jahre spätere schulische Erfolgskriterien substanziell vorhersagen können und zwar auch dann, wenn für weitere leistungsprädiktive Merkmale (z. B. Schulnoten) kontrolliert wird. Wie in den meisten Studien zur Untersuchung der psychometrischen Güte von Testverfahren, wurden diese Ergebnisse für die Gesamtpopulation von SuS einer Schulform ermittelt, ohne dass die Variabilität der Prognosegüte zwischen Schulen berücksichtigt wurde. Es wurde somit nicht hinterfragt, inwiefern die Prognosegüte, die für die Gesamtstichprobe ermittelt wurde, sich auf die Schülerschaft an den Einzelschulen generalisieren lässt.

Diese Frage knüpft an eine sehr lebhafte Diskussion zur Generalisierbarkeit der Prognosegüte von Testverfahren im Rahmen der Personalauswahl an (Murphy, 2000). Inzwischen wird solch empirische Evidenz zur Validitätsgeneralisierung von Tests von aktuellen „Standards“ (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014) explizit eingefordert. In der empirischen Bildungsforschung wurden Fragen zur Validitätsgeneralisierung von Schülerleistungstests in Bezug auf Einzelschulen jedoch noch nie gestellt. Ziel unserer Studie ist es daher, erstmals mit Hilfe eines metaanalytischen Ansatzes (Cheung & Jak, 2016) zu untersuchen, inwiefern die Prognosegüte von

bildungsstandardbasierten Mathematiktests über einen Zeitraum von 2 Jahren über Schulen hinweg generalisierbar ist. Zudem untersuchen wir, wie stark sich Unterschiede in der Prognosegüte zwischen Schulen durch schulspezifische Merkmale (Reliabilität der Tests, Leistungsheterogenität, Leistungsniveau) erklären lassen.

Forschungsstand zur Güte (bildungs-)standardisierter Mathematiktests zur Prognose zukünftiger Testergebnisse und Schulnoten in Mathematik

Die Güte bildungsstandardbasierter Mathematiktests zur Prognose zukünftiger Testergebnisse und Schulnoten wurde bisher vertiefend im Rahmen von zwei Längsschnittstudien (Fuchs & Brunner, 2017; Graf et al., 2016) untersucht. Die Ergebnisse der Regressionsanalysen (mit β als standardisiertem Regressionskoeffizienten) in der Primarstufe zeigten, dass bildungsstandardbasierte Mathematiktests 2 bzw. 3 Jahre später erhobene fachgleiche bildungsstandardbasierte Testergebnisse mit $.56 \leq \beta \leq .66$ prognostizierten; die Prognosegüte für Halbjahresnoten lag bei $.48 \leq \beta \leq .57$ (Fuchs & Brunner, 2017). Diese Werte entsprechen der Güte anderer kommerziell erhältlicher, standardisierter Schulleistungstests wie bspw. dem DEMAT (Fuchs & Brunner, 2017, Elektronisches Supplement [ESM]-1). Die Ergebnisse in der Sekundarstufe für Gymnasiasten zeigten, dass VERA-Mathematiktests in der 8. Klasse die Prüfungsnote in Mathematik für den mittleren Schulabschluss (MSA) mit $\beta = .56$ und die Mathematikjahrgangsnote mit $\beta = .44$ vorhersagen (Graf et al., 2016, Tab. 3). Die ermittelte Prognosegüte war auch hier vergleichbar mit der für kommerziell erhältliche Tests für die Sekundarstufe (Graf et al., 2016, ESM-1). Wichtig ist für die beiden Längsschnittstudien hervorzuheben, dass in beiden ein prognostischer Mehrwert der Tests gegenüber anderen leistungsprädiktiven Merkmalen, insbesondere gegenüber Noten, gezeigt werden konnte. Des Weiteren werden in den Länderberichten für die Vergleichsarbeiten in Thüringen Kennwerte zur Prognosegüte von VERA-Tests berichtet (C. Nachtigall, persönl. Mitteilung, 22.10.2018). So liegt die Spannweite der Prognosegüte von Leistungen im VERA-Mathematiktest in der 6. Klasse für die Leistungen im VERA-Mathematiktest in der 8. Klasse (für die sukzessiven Schülerkohorten der Schuljahre 2009/10 bis 2017/18) zwischen $.68 \leq r \leq .80$. Diese Prognosegüte auf Testleistungen ist vergleichbar mit jener für standardisierte Tests in der Sekundarstufe (ESM-1). Welche bisher – ebenso wie für bildungsstandardbasierte Tests – ausnahmslos für die Gesamtstichproben ermittelt wurden.

Metaanalytischer Ansatz zur Generalisierbarkeit der Befunde zur Testgüte aus Einzelstudien

Fragen zur Generalisierbarkeit der Prognosegüte standardisierter Tests sind integraler Bestandteil der „Standards“ (AERA et al. 2014, S. 18): „An important issue in educational and employment settings is the degree to which validity evidence based on test-criterion relations can be generalized to new situations without further study of validity in that new situation. […]. Thus statistical summaries of past validation studies in similar situations may be useful in estimating test-criterion relationship in a new situation.”.

Ausgelöst durch die Entwicklung der psychometrischen Metaanalyse von Schmidt und Hunter (1977), entwickelte sich v. a. im Bereich der Arbeits- und Organisationspsychologie ein lebhafter Diskurs, wie Fragen zur Generalisierbarkeit der Validität bzw. Güte von Tests über verschiedene Situationen hinweg untersucht werden sollten (Banks & McDaniel, 2012). Dabei haben sich psychometrische Metaanalysen etabliert. Vor Einführung der psychometrischen Metaanalysen wurde i. d. R. für jede neue Anwendungssituation eine empirische Studie zur Validität eines Tests durchgeführt, um die Generalisierbarkeit der Befunde aus vorherigen Validitätsstudien zu untersuchen. Psychometrische Metaanalysen wurden als statistische Methode entwickelt, um unsystematische und systematische Abweichungen der Ergebnisse zwischen Einzelstudien zusammenzufassen und die Gründe für die Variabilität der Ergebnisse über die Berücksichtigung von Moderatoren zu analysieren. Es zeigte sich, dass mit psychometrischen Metaanalysen im Vergleich zu Einzelstudien belastbarere Aussagen zur Generalisierbarkeit der Güte von Tests über variierende Situationen hinweg gewonnen werden konnten (Banks & McDaniel, 2012). Dies ist darauf zurückzuführen, dass innerhalb von Einzelstudien Methodenartefakte nicht ausreichend berücksichtigt werden (können), die jedoch i. d. R. den größten Teil der Variation von Ergebnissen zwischen Einzelstudien hinweg erklären (Schmidt, Hunter & Urry, 1976).

Insgesamt zeigen zahlreiche psychometrische Metaanalysen (v. a. zur Anwendung kognitiver Fähigkeitstests in der Personalauswahl; Pearlman, Schmidt & Hunter, 1980), dass die über Einzelstudien aggregierten Indikatoren der Validität (z. B. Korrelationen zwischen Prädiktoren und Kriterien) i. d. R. höher ausfallen und konsistenter sind, als narrative Forschungsüberblicke dies nahelegen würden (Murphy, 2000). Die damit verbundene Möglichkeit zur Validitätsgeneralisierung wird inzwischen als solide Informationsbasis angesehen, um die Nutzung von standardisierten Tests für praxisrelevante Entscheidungen in

neuen Situationen zu rechtfertigen, ohne dass weitere empirische Studien zur Untersuchung der Validität für diese neue Situation notwendig sind (AERA et al. 2014; Murphy, 2000).

Im starken Gegensatz zur Arbeits- und Organisationspsychologie existieren für den Bildungskontext bisher nur sehr wenige Studien zur Generalisierbarkeit der Prognosegüte von Schulleistungstests. Diese Metaanalysen (Reschly, Busch, Betts, Deno & Long, 2009; Yeo, 2010) untersuchten z. B., inwiefern frühere Lesetestleistungen spätere Lesetestleistungen in den Klassen 1 bis 8 vorhersagen können. Meist war der Prognosezeitraum nicht länger als ein Jahr. Die mittlere Prognosegüte lag bei etwa $r = .70$. In diesen Metaanalysen wurde aber nicht untersucht, ob sich die Prognosegüte der Tests zwischen Schulen unterscheidet.

Im Rahmen psychometrischer Metaanalysen wurden mehrere Methodenartefakte identifiziert, die Validitätskennwerte moderieren können. Dazu gehören u. a. die Messgenauigkeit/Reliabilität sowie die Variationsbreite der Prädiktor- und Kriteriumsmaße (Schmidt & Hunter, 2015): Je höher die Reliabilität und je weniger die Variationsbreite vom Populationswert abweicht, desto geringer sind systematische Verzerrungen von Einzelstudien zur Schätzung der „wahren“ Validität. Bezogen auf Einzelschulen legt dies nahe, dass eine höhere Prognosegüte an Schulen zu erwarten ist, an denen die Messwerte auf Prädiktor- und/oder Kriteriumsseite (a) eine höhere Reliabilität sowie (b) eine höhere Leistungsheterogenität der Schülerschaft aufweisen.

Zudem untersuchen wir die moderierende Wirkung des mittleren Leistungsniveaus der Schule. So gibt es zahlreiche empirische Hinweise für Kompositionseffekte in Bezug auf Testleistungen bzw. Bezugsgruppeneffekte auf Noten. Solche Effekte liegen vor, wenn eine auf Schulebene aggregierte individuelle Leistung von SuS einen eigenständigen Beitrag – über den Beitrag auf Individualebene hinaus – zur Varianzaufklärung einer unabhängigen Variable beiträgt (Harker & Tymms, 2004). Daher erwarten wir, dass sich Kompositions- und Bezugsgruppeneffekte für das mittlere Leistungsniveau an der Schule auf die Prognosegüte zeigen. Jedoch sind differenzielle Effekte zu erwarten, je nachdem welches der beiden untersuchten Leistungskriterien prognostiziert wird.

In Bezug auf die spätere Testleistung gehen wir von positiven Kompositionseffekten aus, d. h. das mittlere Leistungsniveau der Schule hat über die individuelle Leistung hinaus einen positiven Effekt auf die Leistung in Klasse 8 (u. a. Dumont, Neumann, Maaz & Trautwein, 2013). Da die individuelle Leistung positiv mit der späteren Leistung assoziiert ist, impliziert

der positive Kompositionseffekt, dass (ceteri paribus) die Prognosegüte der individuellen Leistung in Klasse 6 mit zunehmendem Leistungsniveau der Schule ansteigen sollte.

In Bezug auf die spätere Note gehen wir von negativen Bezugsgruppeneffekten aus, da Lehrkräfte ihre Notengebung an das Leistungsniveau der Schülerschaft anpassen (u. a. Schmid, Paasch & Katstaller, 2016): Bei gleicher Leistung bekommen SuS in leistungsstarken Lernverbänden schlechtere Noten. Damit hat das mittlere Leistungsniveau der Schule über die individuelle Leistung hinaus einen negativen Effekt auf die Note in Klasse 8. Die individuelle Leistung ist positiv mit der späteren Note assoziiert. Der negative Bezugsgruppeneffekt impliziert, dass (cet. par.) die Prognosegüte der individuellen Leistung in Klasse 6 mit zunehmenden Leistungsniveau der Schule abnehmen sollte.

Forschungsfragen

Bildungsstandardbasierte Tests werden u. a. im Rahmen von VERA bundesweit in Schulen eingesetzt und gehören damit zu den am meisten eingesetzten standardisierten Tests in Deutschland. Jedoch existiert bislang für bildungsstandardbasierte Tests keine einzige Studie im deutschsprachigen Raum zur Validitätsgeneralisierung, obwohl dieser empirische Nachweis von hoher Relevanz für die Abschätzung des diagnostischen Potenzials an den Schulen ist. Denn im Rahmen einer evidenzbasierten Bildungspraxis stellt sich für Lehrkräfte die Frage, inwiefern die Ergebnisse zur Validität bzw. zur Güte bildungsstandardbasierter Tests, die entweder für die Gesamtstichprobe aller Schulen oder für andere Schulen gefunden wurden, auf die eigene Schule zutreffen. Wir gehen deshalb auf Basis eines metaanalytischen Ansatzes drei zentralen Fragen zur Validitätsgeneralisierung bildungsstandardbasierter Tests in der 6. Klasse für die Prognose von Testergebnissen und Noten im Fach Mathematik in der 8. Klasse nach: (1) Wie stark variiert die Güte der Tests zur Prognose zukünftiger Testergebnisse und Noten zwischen Schulen? (2) Wie stark variiert der inkrementelle prognostische Mehrwert der Tests gegenüber Noten zwischen Schulen? (3) Inwiefern wird die Prognosegüte der Tests bzw. der inkrementelle prognostische Mehrwert durch schulspezifische Merkmale moderiert: (a) Reliabilität, (b) Leistungsheterogenität und (c) Leistungsniveau?

Methode

Stichprobe und Prozedur

Zur Untersuchung unserer Forschungsfragen nutzten wir einen Archivlängsschnittdatensatz über 2 Schuljahre. Der Datensatz resultierte durch die regulären VERA-Erhebungen in der 6. und 8. Klasse im Schuljahr 2013/14 bzw. 2015/16 für SuS bzw. Schulen innerhalb eines Bundeslandes[3]. Die öffentlichen Schulen wurden zum Zeitpunkt der 6. Klasse zur Durchführung eines VERA-Tests in einem der drei Fächer Deutsch, Englisch und Mathematik mit ihren SuS verpflichtet. Entschied sich eine Schule für die Durchführung der Mathematiktests, waren alle Lehrkräfte dazu verpflichtet. 81 % hatte sich für die Durchführung der Mathematiktests entschieden. In der 8. Klasse waren alle öffentlichen Schulen zur Durchführung der VERA-Mathematiktests verpflichtet.

Insgesamt resultierte nach Anwendung der Ausschlusskriterien eine positiv selektierte Analysestichprobe mit 11 054 SuS (Details ESM-2). Diese SuS besuchten zwei Schularten: Gymnasien und Schulen, die mehrere Bildungsgänge anbieten; erstere bezeichnen wir nachfolgend mit GY und letztere mit SMBG. Insgesamt lagen für die 78 GY Daten von 4 626 SuS vor: 48 % Mädchen; Alter in der 6. Klasse: M = 12.5 Jahre, SD = 0.52 Jahre; 2 % sprechen zu Hause kein Deutsch. Pro GY lagen im Mittel Daten von 59 SuS vor (Min = 10, Max = 134). Für insgesamt 223 SMBG lagen Daten von 6 428 SuS vor: 47 % Mädchen; Alter in der 6. Klasse: M = 12,7 Jahre, SD = 0.66 Jahre; 2 % sprechen zu Hause kein Deutsch. Pro SMBG lagen im Mittel Daten von 29 SuS vor (Min. = 10, Max. = 115).

Messinstrumente

Bildungsstandardbasierte Mathematiktests

In allen Schulen und Schularten wurden in der 6. bzw. 8. Klasse jeweils derselbe (bildungsstandardbasierte) VERA-Mathematiktest eingesetzt. Alle Tests wurden von Lehrkräften durchgeführt und ausgewertet. Hierzu wurden detaillierte Durchführungs- und Auswertungsmanuale zur Verfügung gestellt.

In der 6. Klassen enthielt der VERA-Mathematiktest 27 Items, die Kompetenzen zu allen 5 Leitideen erfassten: Zahl, Messen, Raum und Form, Funktionaler Zusammenhang sowie Daten und Zufall. Die vorgesehene Bearbeitungszeit betrug 60 Minuten. Die Skalierung des

[3] Zur Vermeidung neuartiger Vergleiche wird das Bundesland nicht genannt.

Tests erfolgte für alle SuS der Analysestichprobe (gemeinsam für beide Schularten) auf der Basis eines eindimensionalen Rasch-Modells (Rasch, 1980) mit dem Statistikprogramm R (Paket: TAM; Robitzsch, Kiefer & Wu, 2018). Die Itemparameter wurden frei auf der Grundlage der vorhandenen Daten geschätzt (ESM-2.2). Separat nach Schulform ergab sich für SMBG/GY eine WLE-Reliabilität = .78/.73.

In der 8. Klasse enthielt der VERA-Mathematiktest 46 Items zu allen 5 Leitideen. Die vorgesehene Bearbeitungszeit betrug 80 Minuten. Der Test wurde vom Institut zur Qualitätsentwicklung im Bildungswesen konzipiert und auf Grundlage großer bundeslandübergreifender Stichproben normiert. Für die Skalierung wurde ein eindimensionales Rasch-Modell genutzt, bei dem die normierten Itemparameter für alle Items herangezogen wurden. Obwohl der Test vorab psychometrisch evaluiert wurde, zeigte sich jedoch, dass für viele Items das Raschmodell nur eine unzureichende Modellpassung aufwies (ESM-2.3). Dennoch wurden die so skalierten Werte für die weiteren Analysen genutzt, um (a) die Generalisierbarkeit auf bundesweit eingesetzte Tests nicht einzuschränken und (b) weil separat nach Schulform zumindest akzeptable WLE-Reliabilitäten resultierten: SMBG/GY = .87/.76.

Schulnoten

In der 6. und 8. Klasse gaben die Lehrkräfte die Halbjahresnoten für die SuS in Mathematik auf der Notenskala von 1 (*sehr gut*) bis 6 (*ungenügend*) an. Diese wurden rekodiert in 6 (*sehr gut*) bis 1 (*ungenügend*), sodass höhere Ausprägungen der Noten bessere Schulleistungen repräsentieren.

Standardisierung der Leistungsdaten an der Gesamtheit der Analysestichprobe

Wir berechneten jeweils z-standardisierte Werte ($M = 0$, $SD = 1$) für die Testleistungen sowie Noten in der 6. und 8. Klasse für die SuS der Analysestichprobe – unabhängig von der besuchten Schulform. Auf diese Weise können wir potenzielle Unterschiede zwischen den Schulformen untersuchen, ohne dass diese durch schulformspezifische Unterschiede in der Streuung der Leistungswerte konfundiert sind (deskriptive Kennwerte ESM-2.4).

Moderatoren

Die Moderatoren zum Leistungsniveau und der Leistungsheterogenität an den Schulen haben wir auf Basis der z-standardisierten Testwerte gebildet. Dafür haben wir den schulspezifischen Mittelwert der Testwerte bzw. die schulspezifische Varianz für alle SuS der jeweiligen Schule für jeden der insgesamt 15 imputierten Datensätze berechnet und über diese den Mittelwert pro Schule ermittelt. Zudem haben wir als weitere Moderatoren die schulspezifische WLE-Reliabilität berechnet (WLErel in TAM). Um Fragen zur relativen Bedeutsamkeit der Moderatoren beantworten zu können (Schmidt & Hunter, 2015), haben wir alle schulspezifischen Kennwerte z-standardisiert ($M = 0$, $SD = 1$) und zwar für die Gesamtheit aller Schulen unserer Analysestichprobe unabhängig von der Schulform (deskriptive Kennwerte ESM-2.5).

Datenanalyse

Umgang mit fehlenden Werten

Für die Analysestichprobe lag der Anteil fehlender Werte pro Variable bei maximal 24 % an SMBG und 17 % an GY für den Test in der 6. Klasse (ESM-2.1). Wir nutzten das multiple Imputationsverfahren MICE (Buuren & Groothuis-Oudshoorn, 2011), um jeweils 15 vollständige Datensätze zu erzeugen. Bei der Imputation der fehlenden Werte berücksichtigten wir das Skalenniveau sowie die hierarchische Datenstruktur (Schulebene). Um die Qualität der imputierten Daten zu verbessern (Collins, Schafer & Kam, 2001), nutzten wir neben den Testwerten und Noten in der 6. und 8. Klasse folgende Hilfsvariablen: Alter, Geschlecht und Angaben zum Sprechen einer nicht deutschen Muttersprache.

Analysemethode

Wir haben unsere Analysen separat für SuS an SMBG und SuS an GY durchgeführt, um potenzielle Schulformunterschiede identifizieren zu können. Für jede Schulform analysierten wir die Daten mit der dreischrittigen metaanalytischen Methode von Cheung und Jak (2016).

Im ersten Schritt haben wir für jede Schulform Teildatensätze für jede Einzelschule erstellt (Cheung & Jak, 2016; Schmidt & Hunter, 2015).

Im zweiten Schritt berechneten wir für Forschungsfrage 1 für jede Einzelschule zwei bivariate Regressionsmodelle (β) mit der Testleistung in der 6. Klasse als Prädiktor und (1) der

Testleistung bzw. (2) der Note in der 8. Klasse als Kriterium. Für Forschungsfrage 2 zur inkrementellen Prognosegüte spezifizierten wir für jede Einzelschule multiple Regressionsmodelle, bei denen wir zusätzlich zur Testleistung für die Note in der 6. Klasse kontrollierten, um adjustierte Regressionskoeffizienten (β_{adj}) zu ermitteln. Weiterhin berechneten wir für Forschungsfrage 2 für jede Einzelschule den Semipartialkorrelationskoeffizienten (*sr*; mit escalc; Viechtbauer, 2010): *sr* und β_{adj} liefern komplementäre Informationen zur inkrementellen Prognosegüte (Aloe & Thompson, 2013). So bildet β_{adj} den inkrementellen Wert der Testleistung in der 6. Klasse auf das Kriterium in der 8. Klasse bei gleichzeitiger Kontrolle der Note ab. Hingegen informiert *sr* über den uniquen korrelativen Zusammenhang der Testleistung in der 6. Klasse mit dem jeweiligen Kriterium, wenn die gemeinsame Varianz zwischen Note und dem jeweiligen Kriterium herausgerechnet wird (Cohen, Cohen, West & Aiken, 2003). Diesen zweiten Analyseschritt führten wir separat mit dem Robust Maximum Likelihood Verfahren (MLR) in Mplus8 (Muthén & Muthén, 1998-2017) mit Hilfe des R-Paketes *MplusAutomation* (Hallquist & Wiley, 2018) über die 15 imputierten Datensätze (mit Ausnahme der Berechnung für *sr*) unter Aggregation der Regeln nach Rubin (1987) durch. Die Berechnung von *sr* basierte auf den aggregierten Kennwerten.

Im dritten Analyseschritt nutzten wir die aggregierten Schulergebnisse, um für jede Schulform Kennwerte mit einer *random-effects* Metaanalyse (Raudenbush, 2009) mit dem R-Paket *metafor* (Viechtbauer, 2010) zu berechnen. Den Empfehlungen von Schmidt und Hunter (2015) folgend berücksichtigten wir dabei ein zentrales Methodenartefakt – den Standardfehler – durch Einbezug der Anzahl von SuS an einer Schule (McDaniel, Rothstein Hirsh, Schmidt, Raju & Hunter, 1986).

Eine wichtige Frage im Kontext der Validitätsgeneralisierung ist die Heterogenität der Effektgrößen. Eine inferenzstatistische Prüfung erfolgte mit dem Q-Test (Hedges & Olkin, 1985): Bei einem nicht signifikanten Ergebnis wird angenommen, dass die Heterogenität der Effektgrößen nicht alleinig auf den Stichprobenfehler zurückgeht. Um das Ausmaß der Heterogenität zu quantifizieren, berechneten wir die Standardabweichung der Kennwerte (τ) sowie *credibility intervals* für ein 95 %-Niveau (CR; Rothstein, 2003). Die CRs geben an, in welchem Bereich 95 % der wahren Kennwerte in einer hypothetischen Population an Schulen liegen, welche aus den realen Schulen besteht, die in unsere Studie einbezogen wurden, sowie Schulen, die zukünftig die Tests einsetzen werden.

Zur Beantwortung von Forschungsfrage 3 führten wir für jede Schulform jeweils getrennt Meta-Regressionen durch. Hierzu berechneten wir standardisierte Regressionsgewichte für die jeweiligen Moderatoren (β_M) und ermittelten, (a) ob diese einen signifikanten Beitrag zur Aufklärung der Heterogenität leisten (s. M: p-Wert) und, (b) ob die Heterogenität der Kennwerte zur Prognosegüte nach Kontrolle der Moderatoren signifikant ist (s. Q: p-Wert).

Alle metaanalytischen Modelle wurden mit dem Restricted Maximum Likelihood Verfahren (REML) geschätzt (Veroniki et al., 2016).

Ergebnisse

Forschungsfrage 1: Prognosegüte für zukünftige Testleistungen und Schulnoten

SMBG

Der Mittelwert zur Prognose zukünftiger Testleistungen lag für SMBG bei β = .66 (Tab. 1A; Kriterium Test). Im Mittel ging also an SMBG eine bessere Testleistung in der 6. Klasse von einer Standardabweichung (= SD) mit einer um .66 SD besseren Testleistung in der 8. Klasse einher. Das CR für die Prognosegüte an SMBG lag zwischen β = .51 bzw. β = .80. Die Güte variierte dabei signifikant zwischen den Einzelschulen (Q: $p < .05$). Ein vergleichbares Ergebnis zeigte sich für die Prognose zukünftiger Noten mit einem mittleren β = .57 (Tab. 1A; Kriterium Note): Im Mittel ging also an SMBG eine bessere Testleistung in der 6. Klasse von einer SD mit einer um .57 SD besseren Note in der 8. Klasse einher. Das CR lag zwischen β = .26 und β = .89; die Prognosegüte variierte signifikant zwischen den Einzelschulen.

Tabelle 1

Metaanalyse zur Vorhersage der Testleistung und der Note in Mathematik in der 8. Klasse durch die Testleistung in der 6. Klasse

	Mittelwert	τ	CR	Q: p
A) SMBG				
Kriterium: Test				
β	.66	.07	[.51;.80]	.01*
β_{adj}[a]	.48	.04	[.40;.56]	.61
sr[a]	.39	.15	[.09;.69]	< .01*
Kriterium: Note				
β	.57	.16	[.26;.89]	< .01*
β_{adj}[a]	.24	.12	[-.01;.48]	< .01*
sr[a]	.16	.18	[-.20;.53]	< .01*
B) GY				
Kriterium: Test				
β	.60	.07	[.46;.74]	.01*
β_{adj}[a]	.43	.07	[.30;.56]	.04*
sr[a]	.33	.05	[.24;.43]	.08
Kriterium: Note				
β	.56	.05	[.45;.67]	.17
β_{adj}[a]	.22	.00[b]	[.18;.26]	.70
sr[a]	.14	< .01	[.12;.16]	.38

Anmerkungen. Metrik der Prädiktor- und Kriteriumvariable (s. Spezifikation im Text). SMBG = Schulen mit mehreren Bildungsgängen, GY = Gymnasien, τ = Standardabweichung des jeweiligen Kennwerts, CR = 95% - credibility inteval, Q:p = p-Wert für Cochran's Q-Test H0: $\tau 2 = 0$, β = Regressionskoeffizient, βadj = adjustierter Regressionskoeffizient, sr = Semipartialkorrelationskoeffizient.

[a] Kontrolle für die Halbjahresnote in Mathematik in der 6. Klasse.

[b] τ = 0 trotz Restricted Maximum Likelihood Verfahren (REML; Veroniki et al., 2016)

* p < .05.

GY

Der Mittelwert zur Prognose der Testleistungen lag für GY bei β = .60 (Tab. 1B; Kriterium Test) mit einem CR von β = .46 bis β = .74. Die Prognosegüte variierte signifikant zwischen den Einzelschulen.

Der Mittelwert zur Prognose zukünftiger Noten lag für GY bei β = .56 (Tab. 1B; Kriterium Note) mit einem CR von β = .45 bzw. β = .67. Die Prognosegüte variierte nicht signifikant zwischen den Einzelschulen.

Forschungsfrage 2: Inkrementeller Mehrwert der Testleistungen bei der Prognose

SMBG

An SMBG lag der inkrementelle Mehrwert zur Prognose von Testleistungen im Mittel bei β_{adj} = .48 (Tab. 1A). Die mittlere Semipartialkorrelation lag bei *sr* = .39. Diese Werte implizieren, dass an SMBG Testleistungen in der 6. Klasse über die Note hinaus einen prognostischen inkrementellen Mehrwert für zukünftige Testleistungen in der 8. Klasse aufweisen. Die CRs lagen bei $.40 \leq \beta_{adj} \leq .56$ bzw. $.09 \leq sr \leq .69$. Dabei variierte β_{adj} nicht signifikant, jedoch *sr* signifikant zwischen den Einzelschulen.

An SMBG lag der inkrementellen Mehrwert zur Prognose von Noten im Mittel bei β_{adj} = .24 bzw. *sr* = .16. Diese Werte implizieren, dass selbst bei Kontrolle der Noten in der 6. Klasse die Testleistungen in der 6. Klasse einen inkrementellen Mehrwert für die Prognose von Noten in der 8. Klasse aufweisen. Die CRs lagen zwischen $-.01 \leq \beta_{adj} \leq .48$ bzw. $-.20 \leq sr \leq .53$. Beide Kennwerte variierten signifikant zwischen den Einzelschulen.

GY

An GY wiesen die Testleistungen einen inkrementellen Mehrwert zur Prognose von Testleistungen auf, der im Mittel bei β_{adj} = .43 bzw. *sr* = .33 lag (Tab. 1B). Die CRs lagen bei $.30 \leq \beta_{adj} \leq .56$ bzw. $.24 \leq sr \leq .43$. Dabei variierte β_{adj} signifikant, jedoch *sr* nicht signifikant zwischen den Einzelschulen.

Die Testleistungen der 6. Klasse leisteten auch einen inkrementellen Mehrwert zur Prognose von Noten, der im Mittel β_{adj} = .22 bzw. *sr* = .14 betrug. Die CRs für die Prognosegüte lagen zwischen $.18 \leq \beta_{adj} \leq .26$ bzw. $.12 \leq sr \leq .16$. Beide Kennwerte variierten nicht signifikant zwischen den Einzelschulen.

Forschungsfrage 3: Moderatoren

SMBG

Die Moderatoranalysen für SMBG zeigten (Tab 2A), dass die schulspezifische Reliabilität in der 8. Klasse einen signifikanten Einfluss auf die Prognosegüte (β) und den inkrementellen Mehrwert (β_{adj}) zur Prognose von Testleistungen sowie die Prognosegüte (β) von Noten in der 8. Klasse hatte (M: $p < .05$). So erhöhte sich bspw. β für die Prognose von Testleistungen an den Einzelschulen um β_M = .12, wenn die Reliabilität der Testleistungen an der Schule um

eine SD anstieg. Durch Berücksichtigung der Reliabilität in der 8. Klasse verringerte sich die Heterogenität von β auf ein nicht signifikantes Niveau (Q: $p > .05$).

Die Prognosegüte (β) und der inkrementelle Mehrwert (β_{adj}) auf Testleistungen sowie auf Noten konnte durch die Leistungsheterogenität in der 8. Klasse signifikant erklärt werden. Damit reduzierte sich die Heterogenität für die Prognosegüte von Testleistungen auf ein nicht signifikantes Ausmaß; die Heterogenität zur Prognose der Noten blieb signifikant.

Entgegen unseren Erwartungen konnte das schulspezifische Leistungsniveau keinen signifikanten Beitrag zur Aufklärung der Heterogenität für die Prognosegüte und den inkrementellen Mehrwert leisten – weder für die Prognose der Testleistung noch für die der Note.

Tabelle 2

Meta-Regressionen zur Erklärung von Unterschieden zwischen Schulen in der Prognosegüte für Testleistungen in der 6. Klasse

	Rel 6. Klasse			Rel 8. Klasse			LH 6. Klasse			LH 8. Klasse			LN 6. Klasse			LN 8. Klasse		
	β_M	M:p	Q:p	β_M	M:p	Q:p	β_M	M:p	Q:p	β_M	M:p	Q:p	β_M	M:p	Q:p	β_M	M:p	Q:p
A) SMBG																		
Kriterium Test																		
β	.02	.18	.01*	.12	<.01*	1.00	.01	.34	.01*	.10	<.01*	1.00	.01	.73	.02*	.02	.22	.01*
β_{adj}[a]	.01	.71	.60	.07	<.01*	.99	.01	.65	.61	.06	<.01*	.99	.01	.70	.63	.01	.52	.61
sr[a]	.01	.41	<.01*	<.01	.94	<.01*	.01	.44	<.01*	<.00	.96	<.01*	.01	.67	<.01*	.01	.58	<.01*
Kriterium: Note																		
β	<.01	.99	<.01*	.08	<.01*	<.01*	.01	.74	<.01*	.07	<.01*	<.01*	-.05	.10	<.01*	-.04	.14	<.01*
β_{adj}[a]	<.01	.95	<.01*	.04	.08	.01*	.01	.74	<.01*	.04	.03*	.01*	.01	.82	<.01*	-.02	.54	<.01*
sr[a]	.01	.59	<.01*	.03	.17	<.01*	.01	.55	<.01*	.03	.07	<.01*	.02	.52	<.01*	<.00	.89	<.01*
B) GY																		
Kriterium Test																		
β	.04	.11	.02*	.09	<.01*	.76	<.01	.89	.01*	.13	<.01*	.97	-.10	<.01*	.07	-.06	.12	.02*
β_{adj}[a]	.04	.18	<.05*	.09	<.01*	.53	-.01	.57	.03*	.12	<.01*	.57	-.13	<.01*	.11	-.07	.06	.06
sr[a]	.04	.04*	.15	.04	<.01*	.22	.02	.38	.09	.04	.04*	.11	-.06	.06	.10	-.05	.08	.11
Kriterium: Note																		
β	-.03	.37	.16	.02	.34	.15	-.02	.40	.15	.03	.31	.16	.04	.41	.17	<.00	.93	.15
β_{adj}[a]	-.03	.48	.67	.02	.27	.69	-.04	.14	.70	<.00	.90	.66	-.03	.45	.67	-.05	.25	.69
sr[a]	-.01	.73	.35	.01	.73	.35	-.01	.61	.36	<.00	.89	.35	.01	.85	.35	-.01	.71	.35

Anmerkungen. Metrik der Prädiktor- und Kriteriumvariable (s. Spezifikation im Text). Rel = Reliabilität, LH = Leistungsheterogenität, LN = Leistungsniveau. SMBG = Schulen mit mehreren Bildungsgängen, GY = Gymnasien, β_M: standardisiertes Regressionsgewicht für den jeweiligen Moderator, M:p = p-Wert für Omnibus-Test H0: $\beta_M = 0$; Q:p = p-Wert für Cochran's Q-Test H0: $\tau^2 = 0$; Kennwerte aus Regressionsmodellen pro Schule: β = Regressionskoeffizient, β_{adj} = adjustierter Regressionskoeffizient, sr = Semipartialkorrelationskoeffizient.

[a] Kontrolle für die Halbjahresnote in Mathematik in der 6. Klasse.

* $p < .05$.

GY

Die schulspezifische Reliabilität in der 8. Klasse hatte einen statistisch signifikanten Einfluss auf die Höhe der Prognosegüte (β) und den inkrementellen Mehrwert ($\beta_{adj,}$ *sr*) zur Prognose der Testleistungen. Bei Kontrolle der Reliabilität reduzierte sich die Heterogenität dieser Kennwerte auf ein nicht signifikantes Ausmaß. Dieses Ergebnismuster zeigte sich ebenso für die schulspezifische Reliabilität in der 6. Klasse für den inkrementellen Mehrwert (*sr*) zur Prognose der Testleistungen.

Ein vergleichbares Ergebnismuster zeigte sich auch für die schulspezifische Leistungsheterogenität in der 8. Klasse für die Prognosegüte (β) und den inkrementellen Mehrwert ($\beta_{adj,}$ *sr*) zur Vorhersage der Testleistungen. Bei Kontrolle der Leistungsheterogenität reduzierte sich die Heterogenität dieser Kennwerte auf ein nicht signifikantes Ausmaß.

Einen signifikanten Beitrag konnte das schulspezifische Leistungsniveau in der 6. Klasse zur Aufklärung der Heterogenität für die Prognosegüte (β) und den inkrementellen Mehrwert (β_{adj}) für die Prognose der Testleistung leisten. Entgegen unserer Erwartungen ermittelten wir ein negatives Regressionsgewicht in Bezug auf die Prognosegüte für Testleistungen: An GY mit höherem mittleren Leistungsniveau, resultierte eine niedrigere Prognosegüte der Tests auf spätere Testleistungen.

Diskussion

Im Bereich der Arbeits- und Organisationspsychologie sind psychometrische Metaanalysen inzwischen ein etabliertes Verfahren zur Prüfung der Validitätsgeneralisierung standardisierter Tests. Für bildungsstandardbasierte Tests existierte bislang keine einzige Studie zur Validitätsgeneralisierung, obwohl solche im Rahmen von VERA-Tests bundesweit an allen Schulen eingesetzt werden und Lehrkräfte für eine evidenzbasierte diagnostische Praxis wissen müssen, inwiefern VERA-Tests das Potenzial haben, auch an ihrer Schule die Leistungsdiagnostik zu verbessern. Wir haben im Rahmen dieses Artikels deshalb die Validitätsgeneralisierung der Prognosegüte bildungsstandardbasierter Tests in der 6. Klasse für Testleistungen und Noten in der 8. Klasse für das Fach Mathematik untersucht.

Prognosegüte bildungsstandardbasierter Tests

Gemittelt über alle Einzelschulen an SMBG bzw. alle GY zeigte sich jeweils, dass der (bildungsstandardbasierte) VERA-Mathematiktest in der 6. Klasse substanziell zukünftige VERA-Mathematiktestleistungen und Mathematiknoten in der 8. Klasse vorhersagen kann. Die Stärke der Prognosegüte ist dabei vergleichbar mit den Ergebnissen, die in bisherigen Studien für bildungsstandardbasierte Tests sowie standardisierte Schulleistungstests ermittelt wurden. Es zeigte sich dabei, dass sowohl für SMBG als auch für GY die Prognosegüte auf zukünftige Testleistungen signifikant zwischen den Einzelschulen variierte. An SMBG variierte auch die Prognosegüte auf zukünftige Noten signifikant zwischen Einzelschulen. Trotz dieser eingeschränkten Validitätsgeneralisierung weisen jedoch die Untergrenzen der Verteilungen (der 95 %-CRs) daraufhin, dass an allen Schulen beider Schulformen mit bildungsstandardbasierten Tests eine bedeutsame Vorhersage auf spätere Testleistungen wie auch Noten möglich ist. So liegt trotz der Heterogenität zwischen Schulen für mind. 93 % der SMBG und 100 % der GY die Prognosegüte der Tests auf Testleistungen und Noten über $\beta = 0$ (ESM-3). Hervorzuheben ist, dass die Prognosegüte auf Noten nicht signifikant zwischen GY variierte, was darauf hinweist, dass für GY die Ergebnisse zu diesem Aspekt der Prognosegüte von VERA-Mathematiktests generalisierbar sind.

Inkrementeller prognostischer Mehrwert bildungsstandardbasierter Tests gegenüber Schulnoten

Gemittelt über alle Einzelschulen an SMBG bzw. alle GY zeigte sich jeweils, dass der VERA-Mathematiktest in der 6. Klasse einen inkrementellen Mehrwert gegenüber Mathematiknoten zur Prognose von VERA-Mathematiktestleistungen und für die Mathematiknoten in der 8. Klasse aufwies. Dies steht im Einklang mit den Befunden von Graf und Kollegen (2016) sowie Fuchs und Brunner (2017), in denen ebenso (auf Grundlage anderer Stichproben) ein inkrementeller Mehrwert auf spätere Noten bzw. Testleistungen für bildungsstandardbasierte Tests nachgewiesen wurde.

Der inkrementelle Mehrwert zur Prognose von Testleistungen variiert z. T. zwischen SMBG (für *sr*) und GY (für β_{adj}) signifikant – jedoch liegt der Mehrwert für 81 % der SMBG bzw. 100 % der GY über $sr/\beta_{adj} = 0$ (ESM-3). Auch der inkrementelle Mehrwert zur Prognose von zukünftigen Noten variiert signifikant zwischen SMBG (β_{adj}, *sr*). Trotzdem liegt für mindestens 60 % der SMBG ein inkrementeller Mehrwert vor (ESM-3.1). Hingegen besteht

für GY keine Heterogenität zwischen den Schulen in Bezug auf die diesbezüglichen Kennwerte. Dies weist darauf hin, dass an allen GY Tests in der 6. Klasse substanziell und über Noten hinaus, die Testleistungen bzw. Noten in der 8. Klasse vorhersagen können.

Moderatoren zur Erklärung der Heterogenität der Prognosegüte und des inkrementellen prognostischen Mehrwerts

Sowohl an SMBG als auch GY ließ sich die Heterogenität in der Prognosegüte und dem inkrementellen prognostischen Mehrwert zwischen den Einzelschulen z. T. durch schulspezifische Merkmale erklären. An beiden Schulformen konnten die schulspezifische Reliabilität und die Leistungsheterogenität in der 8. Klasse den mit bedeutsamsten Beitrag zur Erklärung von Unterschieden leisten: Erwartungskonform ließen sich für Schulen mit höherer mittlerer Reliabilität und Leistungsheterogenität eine höhere Prognosegüte und ein höherer inkrementeller Mehrwert zur Prognose beider zukünftiger Leistungskriterien aufzeigen.

An SMBG zeigte sich weiterhin, dass sich die Heterogenität zwischen den Einzelschulen in der Prognosegüte von Testleistungen bei Berücksichtigung der Reliabilität bzw. der Leistungsheterogenität in der 8. Klasse auf ein nicht signifikantes Niveau reduzieren ließ. Schulunterschiede in der Prognosegüte auf Noten sowie für die meisten Kennwerte des inkrementellen Mehrwerts blieben sowohl für die Testleistung als auch für die Note in der 8. Klasse bestehen.

Auch an GY zeigte sich, dass sich die Heterogenität der Prognosegüte zwischen Schulen bei Berücksichtigung des Leistungsniveaus in der 6. Klasse, der Leistungsheterogenität in der 8. Klasse bzw. der Reliabilität in der 8. Klasse (und für einen Kennwert der Reliabilität in der 6. Klasse) jeweils auf ein nicht signifikantes Niveau reduzieren ließ.

Für das mittlere Leistungsniveau an GY und SMBG zeigte sich jeweils ein erwartungswidriges Befundmuster. Für SMBG konnte weder das Leistungsniveau in der 6. noch in der 8. Klasse die Heterogenität in der Prognosegüte signifikant erklären. An GY zeigten sich sogar negative signifikante Zusammenhänge des mittleren Leistungsniveaus in der 6. Klasse mit der Prognosegüte auf Testleistungen. Ein Grund für die erwartungswidrigen Ergebnisse könnte sein, dass die Kontext- und Bezugsgruppeneffekte auf Schulebene zu schwach sind und sich deutlich ausgeprägter auf Klassenebene zeigen (Ewijk & Sleegers, 2010). Die negativen Effekte für GY zur Erklärung der Heterogenität der Prognosegüte für

Testleistungen könnten darüber hinaus zustande gekommen sein, weil hier der Kompositionseffekt durch Methodenartefakte überlagert wurde. So ist im Kontext von raschskalierten Tests gut dokumentiert, dass die Reliabilität zunimmt, je besser die Itemschwierigkeiten der Tests zum Leistungsniveau der Personen passen (Furr, 2018). Bildungsstandardbasierte Tests sind für SuS im mittleren bis niedrigeren Leistungsniveau optimiert und somit weniger reliabel für (sehr) leistungsstarke SuS. Folglich sollte die Reliabilität an GY mit leistungsstarker Schülerschaft geringer ausfallen (s. r = -.26/-.27 zwischen schulspezifischer Reliabilität und schulspezifischem Leistungsniveau in der 6./8. Klasse in ESM-2.5; s. a. Details in ESM-4). Eventuell wird der negative Effekt dadurch verstärkt, dass die Tests an GY mit einer leistungsstarken Schülerschaft im oberen Leistungsbereich nicht mehr differenzierten. Dieser Deckeneffekt reduziert die Leistungsheterogenität und damit (wie die Moderatoranalysen zeigen) die Prognosegüte. Insgesamt kann selbst bei Vorliegen eines Kompositionseffekts dies dazu führen, dass an leistungsstarken GY die Prognosegüte geringer ausfällt als an weniger leistungsstarken GY.

Grenzen

Die Generalisierbarkeit unserer Ergebnisse könnte durch mehrere Grenzen eingeschränkt sein. Erstens, zur Erfassung aller fachspezifischen Kompetenzen werden jedes Schuljahr neu zusammengestellte VERA-Tests eingesetzt. Es gibt dabei keine Überlappung der eingesetzten Items zwischen VERA-Tests verschiedener Schuljahre. Es ist eine empirisch offene Frage, inwiefern sich die vorliegenden Befunde, die in Mathematik auf einer bestimmten Itemauswahl basieren, auf andere Itemkombinationen oder sogar Fächer generalisieren lassen.

Zweitens liegt in unserer Studie eine positiv selektierte Stichprobe vor, da SuS mit schwächeren schulischen Leistungen aus der Analysestichprobe ausgeschlossen wurden (bspw. Graf et al., 2016). Dies könnte die Varianzen von Tests und Noten insgesamt einschränkt haben. Daher betrachten wir die hier berichteten Mittelwerte für die Kennwerte als eine konservative Schätzuntergrenze.

Drittens können wir nicht abschließend klären, ob der Längsschnittdatensatz repräsentativ für alle Schulen des Bundeslandes war. Wir können nicht ausschließen, dass die Freiwilligkeit zur Teilnahme zum Zeitpunkt der 6. Klasse möglicherweise mit Schulmerkmalen zusammenhängt, die Einfluss auf die Befunde nehmen können. Selektivitätsanalysen mit Hilfe von *funnel plots* geben hierzu jedoch keinen Hinweis, da sie insgesamt zeigen, dass

Schulen mit niedriger und hoher Prognosegüte nicht über- bzw. unterrepräsentiert sind (ESM-5).

Viertens, in der vorliegenden Arbeit haben wir uns dazu entschieden, ein metaanalytisches Vorgehen von Cheung und Jak (2016) zur Untersuchung unserer Forschungsfragen zu verwenden, da dieser die größtmögliche Flexibilität bei der Berechnung statistischer Kennwerte – insbesondere des inkrementellen Mehrwerts – bot. Wenn mit Hilfe von Mehrebenenanalysen vergleichbare Kennwerte berechnet werden können, zeigen sich keine bedeutsamen Unterschiede zu denen, die mit einem solchen metaanalytischen Vorgehen ermittelt werden (Cheung & Jak, 2016). Dies bestätigen auch Mehrebenbenenalysen, die wir für die vorliegenden Daten durchgeführt haben (ESM-6): Auch hier gab es nur marginale Unterschiede zwischen den statistischen Methoden. Dennoch sollten zukünftige Studien klarer die Bedingungen herausarbeiten, unter denen es zu Abweichungen zwischen den statistischen Verfahren kommen kann bzw. die Grenzen des hier verwendeten metaanalytischen Verfahrens evaluieren (bspw. Umgang mit fehlenden Werten).

Schlussfolgerung

Unsere Studie liefert erstmalig empirische Evidenz dafür, dass sowohl an der großen Mehrzahl der GY und den Einzelschulen der SMBG bildungsstandardbasierte Tests in der 6. Klasse Testleistungen und Noten im Fach Mathematik in der 8. Klasse substanziell vorhersagen können und zwar auch dann, wenn die Noten in der 6. Klasse kontrolliert werden. An den meisten Schulen liefern also bildungsstandardbasierte Tests in der 6. Klasse einen substanziellen Informationszugewinn für die Leistungsdiagnostik über Zeugnisnoten hinaus. Dies ist ein bemerkenswertes Ergebnis, wenn man berücksichtigt, dass es sich beim Testergebnis lediglich um eine einmalige Momentaufnahme im Vergleich zur untersuchten Zeugnisnote handelt, die sich aus zahlreichen Leistungsmessungen zusammensetzt und dadurch reliabler sein könnte. Eine solch umfangreiche empirische Befundlage zur prognostischen (inkrementellen) Validität liegt bisher unseres Wissens in diesem Ausmaß für keinen anderen deutschsprachigen, standardisierten Schulleistungstests vor. Zum Teil besteht jedoch ein substanzielles Ausmaß an Heterogenität in der Prognosegüte zwischen den Schulen, was die Validitätsgeneralisierung einschränkt. Jedoch ist es möglich, diese Einschränkung zumindest teilweise zu kompensieren, wenn die Reliabilität der Tests erhöht wird, indem das Schwierigkeitsniveau der Testitems besser an das Leistungsniveau der SuS

angepasst wird. Unsere Ergebnisse unterfüttern somit empirisch, aktuelle Bestrebungen der KMK, die eine Flexibilisierung und Modularisierung der Testdurchführung anzielt, um durch passgenauere Testmaterialien (bspw. zum Leistungsniveau der SuS) den diagnostischen Erkenntnisgewinn an den Schulen zu fördern (KMK, 2012).

Literatur

Aloe, A. M. & Thompson, C. G. (2013). The synthesis of partial effect sizes. *Journal of the Society for Social Work and Research, 4* (4), 390–405.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Banks, G. C. & McDaniel, M. A. (2012). Meta-analysis as a validity summary tool. In N. Schmitt (Hrsg.), *The Oxford handbook of personnel assessment and selection* (S. 156–175). New York, NY: Oxford University Press.

Buuren, S. van & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal Of Statistical Software, 45* (3), 1–67.

Cheung, M. W.-L. & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 1-19.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3. Auflage). Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6* (4), 330–351.

Dumont, H., Neumann, M., Maaz, K. & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen: Internationale und nationale Befunde. *Psychologie in Erziehung und Unterricht, 60* (3), 163–183.

Ewijk, R. van & Sleegers, P. (2010). Peer ethnicity and achievement: a meta-analysis into the compositional effect. *School Effectiveness and School Improvement, 21* (3), 237–265.

Fuchs, G. & Brunner, M. (2017). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie, 31* (1), 27–39.

Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Los Angeles, CA: SAGE.

Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie, 30* (4), 201–211.

Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus structural equation modeling, 1–18.

Harker, R. & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15* (2), 177–199.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Aufl.). Weinheim: Beltz Verlag.

Kultusministerkonferenz. (2004). *Bildungsstandards der Kultusministerkonferenz: Erläuterungen zur Konzeption und Entwicklung (Am 16.12.2004 von der Kultusministerkonferenz zustimmend zur Kenntnis genommen)*. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Konzeption-Entwicklung.pdf

Kultusministerkonferenz. (2012). *Vereinbarung zur Weiterentwicklung von VERA (Beschluss der Kultusministerkonferenz vom 08.03.2012 i.d. F. vom 10.03.2018)*. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Kultusministerkonferenz (Hrsg.). (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Berlin: Kluwer. Verfügbar unter: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

McDaniel, M. A., Rothstein Hirsh, H., Schmidt, F. L., Raju, N. S. & Hunter, J. E. (1986). Interpreting the results of meta-analystic research: A comment on Schmitt, Gooding, Noe, and Kirsch (1984). *Personnel Psychology, 39* (1), 141–148.

Murphy, K. R. (2000). Impact of assessments of validity generalization and situational specificity on the science and practice of personnel selection. *International Journal of Selection and Assessment*, *8* (4), 194–206.

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Pearlman, K., Schmidt, F. L. & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, *65* (4), 373–406.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Edition.). Chicago, IL: University of Chicago Press.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L.V. Hedges & J.C. Valentine (Hrsg.), *The handbook of research synthesis and meta-analysis* (2nd ed., S. 295–315). New York, NY: Russell Sage Foundation.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L. & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47* (6), 427–469.

Robitzsch, A., Kiefer, T. & Wu, M. (2018). *TAM: Test analysis modules*. Verfügbar unter: https://CRAN.R-project.org/package=TAM

Rothstein, H. R. (2003). Progress is our most important product: Contributions of validity generalization and meta-analysis to the development and communication of knowledge in I/0 psychology. In K.R. Murphy (Hrsg.), *Validity generalization: A critical review* (S. 115–154). Mahwah, NJ: Lawrence Erlbaum.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Schmid, C., Paasch, D. & Katstaller, M. (2016). Kompositionseffekte bei der Notenvergabe in Mathematik auf der 4. Schulstufe der österreichischen Volksschule. *Zeitschrift für Bildungsforschung, 6* (3), 265–283.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62* (5), 529–540.

Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis: correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: SAGE.

Schmidt, F. L., Hunter, J. E. & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61* (4), 473–485.

Statistisches Bundesamt. (2017). *Bildung und Kultur: Allgemeinbildende Schulen.* Verfügbar unter: https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/AllgemeinbildendeSchulen2110100177004.pdf?__blob=publicationFile

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G. et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7* (1), 55–79.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36* (3), 1–48.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education*, *31* (6), 412–422.

Elektronische Supplemente (ESM) – Studie II

ESM-1.

Forschungsstand zur Prognosegüte standardisierter Leistungstests auf spätere Testleistungen innerhalb der Sekundarstufe.

ESM-2.

Ergänzungen zur Methode.

ESM-3.

Veranschaulichung der Kennwerteverteilungen für die Einzelschulen auf Basis der CR.

ESM-4.

Ergänzende Erläuterungen für den Zusammenhang zwischen Reliabilität und Leistungsniveau auf Schüler- und Schulebene.

ESM-5.

Funnel plots für die jeweiligen Kennwerte der Prognosegüte.

ESM-6.

Methodenvergleich zur Ermittlung der Prognosegüte.

Tabelle ESM-1: Forschungsstand zur prognostischen Güte standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten

Test	Prognosedauer	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Mathematiktest und Mathematiktest (kommerzielle Tests)							
MAESTRA 5-6+	1 Monat	5.	DEMAT 5+	$r = .15^{a}$			Lingel, Götz, Artelt & Schneider, 2014
MAESTRA 5-6+	1 Monat	6.	DEMAT 6+	$r = .36^{a}$			Lingel et al., 2014
MAESTRA 5-6+	1 Monat	k. A.	MAESTRA 5-6+	$r = .70^{a}$			Lingel et al., 2014
DEMAT 5+	1 Monat	5.	DEMAT 5+	$r = .85^{a}$			Götz, Lingel & Schneider, 2013
DEMAT 6+	1 Monat	6.	DEMAT 6+	$r = .91^{a}$			Götz, Lingel & Schneider, 2013a
BST	2-2,5 Monate	8./9.	BST	HS, RS: $r = .93$			Birkel, Schein & Schumann, 2002
KRW 9	2 Monate	9.	KRW 9	$r = .77^{a}$			Schmidt, Ennemoser & Krajewski, 2013
TeMaTex	2-4 Monate	k. A.	TeMaTex	BK: $r = .78$			Jordan & Stein, 2011
TeMaDi	2-4 Monate	5.	TeMaDi	GS: $r = .71$			Stecken & Stein, 2015

(Tabelle ESM-1 wird fortgesetzt)

Tabelle ESM-1 (Fortsetzung): Forschungsstand zur prognostischen Güte standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten

Test	Prognosedauer	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Mathematiktest und Mathematiktest (nicht kommerzielle Tests)							
TIMSS 1993/94	1 Jahr	7.- 8.	TIMSS 1994/95	Test-Dimension 1: Gesamt: r = .65 GY: r = .54 RS: r = .46 HS: r = .49 Test-Dimension 2: Gesamt: r = .66 GY: r = .46 RS: r = .46 HS: r =.58			Becker, Lüdtke, Trautwein & Baumert, 2006
HarmoS-Test für Ende Klasse 6 2010	≈ 3 Jahre	Anfang 7.- Ende 9.	HarmoS-Test für Ende Klasse 9 2013	r = .71			Wälti, 2014

(Tabelle ESM-1 wird fortgesetzt)

Tabelle ESM-1 (Fortsetzung): Forschungsstand zur prognostischen Güte standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten

Test	Prognosedauer	Prognosezeitpunkt (Klassenstufe)	Kriterium	Prognosegüte	Zusätzliche Prädiktoren	Prognostischer Mehrwert	Quelle
Zusammenhang zwischen Mathematiktest und Mathematiknote (kommerzielle und nicht kommerzielle Tests)							
HarmoS-Test für Ende Klasse 6 2010	≈ 3 Jahre	Anfang 7.- Ende 9.	Note	r = .50[b]			Wälti, 2014

Anmerkungen. MAESTRA 5-6+ = Mathematisches Strategiewissen für fünfte und sechste Klassen. DEMAT 5+ = Deutscher Mathematiktest für fünfte Klassen. DEMAT 6+ = Deutscher Mathematiktest für sechste Klassen. BST = Bausteine-Test. KRW 9 = Konventions- und Regelwissen (Ergänzungstest zu DEMAT 9). DEMAT 9 = Deutscher Mathematiktest für neunte Klassen. TeMatex = Test zum mathematischen Textverständnis. TeMaDi = Test zum mathematischen Diagrammverständnis. TIMSS = Trends in International Mathematics and Science Study. HarmoS-Test für Ende Klasse 6. HS = Hauptschule. RS = Realschule. BK = Berufskolleg. GS = Gesamtschule. GY = Gymnasium. r = Korrelationskoeffizient. r_{adj} = adjustierter Korrelationskoeffizient.

[a] basiert auf Teilstichprobe (für Spezifikation siehe Manual des jeweiligen Tests). [b] Transformation der Schulnote in Abhängigkeit der Schulform.

ESM-1: Literatur

Baumert, J., Roeder, P. M., Gruehn, S., Heyn, S., Köller, O., Rimmele, R. et al. (1996). Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU). In K.-P. Treumann, G. Neubauer, R. Möller & J. Abel (Hrsg.), *Methoden und Anwendung empirischer pädagogischer Forschung* (S. 170–180). Münster: Waxmann.

Becker, M., Lüdtke, O., Trautwein, U. & Baumert, J. (2006). Leistungszuwachs in Mathematik: Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? *Zeitschrift für Pädagogische Psychologie*, *20* (4), 233–242. https://doi.org/10.1024/1010-0652.20.4.233

Birkel, P., Schein, S. A. & Schumann, H. (2002). *BST: Bausteine-Test*. Göttingen: Hogrefe.

Götz, L., Lingel, K. & Schneider, W. (2013a). *DEMAT 6+: Deutscher Mathematiktest für sechste Klassen*. (M. Hasselhorn, W. Schneider & U. Trautwein, Hrsg.). Göttingen: Hogrefe.

Götz, L., Lingel, K. & Schneider, W. (2013). *DEMAT 5+: Deutscher Mathematiktest für fünfte Klassen*. (M. Hasselhorn, W. Schneider & U. Trautwein, Hrsg.). Göttingen: Hogrefe.

Jordan, R. & Stein, M. (2011). *TeMaTex: Test zum mathematischen Textverständnis*. Münster: Verlag für Wissenschaftliche Texte und Medien.

Lingel, K., Götz, L., Artelt, C. & Schneider, W. (2014). *MAESTRA 5-6+: Mathematisches Strategiewissen für fünfte und sechste Klassen*. (M. Hasselhorn, W. Schneider & U. Trautwein, Hrsg.). Göttingen: Hogrefe.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M. et al. (Hrsg.). (2006). *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster: Waxmann.

Schmidt, S., Ennemoser, M. & Krajewski, K. (2013). *DEMAT 9: Deutscher Mathematiktest für neunte Klassen mit Ergänzungstest Konventions- und Regelwissen*. (M. Hasselhorn, W. Schneider & U. Trautwein, Hrsg.). Göttingen: Hogrefe.

Stecken, T. & Stein, M. (2015). *TeMaDi: Test zum mathematischen Diagrammverständnis*. Münster: Verlag für Wissenschaftliche Texte und Medien.

Wälti, B. S. (2014). *Alternative Leistungsbewertung in der Mathematik*. Technische Universität Darmstadt.

ESM-2: Ergänzungen zur Methode

Stichprobe und Prozedur: Definition der Analysestichprobe

Der Analysedatensatz für das Fach Mathematik der vorliegenden Studie wurde rückwirkend von den Archivdaten in der 8. Klasse erstellt. Folglich gibt es keinen systematischen Stichprobenausfall, der üblicherweise in Längsschnittstudien entsteht, in denen der Datensatz ausgehend von der ersten Erhebung prospektiv erstellt wird. Im Archivdatensatz für das Fach Mathematik lagen insgesamt Daten von 12258 SuS aus 324 öffentlichen Schulen vor (exklusive Förderschulen u. a. aufgrund der geringen Schulanzahl). Zur Definition der Analysestichprobe haben wir n = 348 SuS mit sonderpädagogischem Förderbedarf (Hören, Sehen, körperliche und motorische Entwicklung, Lernen und geistige Entwicklung, emotionale und soziale Entwicklung) sowie Jugendliche mit besonderen Lernschwächen für Schreiben, Verhalten oder Rechnen ausgeschlossen, da die VERA-Tests – nach bisherigem Kenntnisstand – für diese Kinder nur bedingt geeignet sein dürften (Pohl, Südkamp, Hardt, Carstensen & Weinert, 2016; Südkamp, Pohl & Weinert, 2015). SuS, die die Schule im Verlauf des Erhebungszeitraumes gewechselt hatten, wurden ebenfalls von den Analysen ausgeschlossen, da für diese keine schulspezifischen Analysen möglich sind (n = 797). Zudem wurden Schulen mit weniger als 10 SuS zum Zeitpunkt der 6. Klasse von den Analysen ausgeschlossen, um robuste Schätzungen für die schulspezifischen statistischen Koeffizienten zu erhalten (n = 8 Schulen; n = 59 Jugendliche; Hox, 2010).

Um einschätzen zu können, ob die Jugendlichen in der Analysestichprobe vergleichbar mit den Jugendlichen waren, die von den Analysen ausgeschlossen wurden, haben wir mehrere Merkmale dieser beiden Schülergruppen schulformüberreifend untersucht (ESM-2.1). Es zeigten sich im Vergleich zu allen ausgeschlossenen SuS die größten Unterschiede ($-0.29 \leq d \leq -.0.51$) in den Leistungskriterien zugunsten der Analysestichprobe. Insgesamt weist diese Analyse darauf hin, dass mit der Analysestichprobe eine positiv selektierte Schülergruppe untersucht wurde, in der (sicherlich auch aufgrund des Ausschlusses von SuS mit sonderpädagogischen Förderbedarf) leistungsschwächere SuS unterrepräsentiert sind.

Tabelle ESM-2.1: Analysen zum Stichprobenausfall und Umfang fehlender Werte für die Analysestichprobe

	von Analysen ausgeschlossene SuS (*N* = 1204)			in Analysen einbezogene SuS (*N* = 11054)				Missinganteil (%) für die in Analysen einbezogene SuS separat pro Schulform	
	N	*M*	*SD*	*N*	*M*	*SD*	*d*	*SMBG*	*GY*
Testleistung									
6. Klasse	828	-0.63	1.22	8868	0.01	1.23	-0.51	17	24
8. Klasse	1078	436	109	10068	497	115	-0.53	10	7
Schulnoten									
6. Klasse (r)[a]	929	3.83	0.99	9216	4.30	0.92	-0.50	13	21
8. Klasse (r)	1191	3.69	1.05	10997	4.00	1.05	-0.29	1	0.5
Merkmale in der 6. Klasse									
Alter (in Jahren)	1204	12.76	0.70	11054	12.59	0.61	0.27	0	0
Anteil der Mädchen	1204	0.43	0.49	11054	0.47	0.50	-0.10	0	0
Anteil der Kinder mit Deutsch als Mutter-sprache	1189	0.02	0.13	10907	0.02	0.12	0.02	1	1

Anmerkungen. SuS = Schülerinnen und Schüler, N = Stichprobengröße, M = Mittelwert, SD = Standardabweichung, d = Cohen's d mit gepoolter Standardabweichung, SMBG = Schulen mit mehreren Bildungsgängen, GY = Gymnasien. (r) = rekodiert (6 = sehr gut – 1= ungenügend).
a zur Bestimmung der Leistungskennwerte für die ausgeschlossenen Schülerinnen und Schüler wurden die Itemparameter zugrunde gelegt, welche aus der freien Skalierung für die Analysestichprobe gewonnen wurden.

Messinstrumente: Infit/Outfit der bildungsstandardbasierten Mathematiktestitems

Bildungsstandardbasierter Test in der 6. Klasse

In der 6. Klassenstufe enthielt der VERA-Mathematiktest 27 Items, die Kompetenzen zu allen fünf Leitideen erfassten: Zahl, Messen, Raum und Form, Funktionaler Zusammenhang sowie Daten und Zufall. Die vorgesehene Bearbeitungszeit betrug 60 Minuten. Der Test wurde von einem Länderverbundprojekt konzipiert und entwickelt. Die Skalierung des Kompetenztests erfolgte für alle Schülerinnen und Schüler der Analysestichprobe (gemeinsam für beide Schularten) auf der Basis eines eindimensionalen Rasch-Modells (Rasch, 1980) für dichotome Daten mit dem Statistikprogramm R (Paket: TAM; Robitzsch, Kiefer & Wu, 2018). Die Itemparameter wurden frei auf der Grundlage der vorhandenen Daten geschätzt. Die Outfitmaße lagen für 23 Items zwischen 0.8 und 1.2 (4 Items = 1.3) und in Bezug auf die Infitmaße für alle Items zwischen 0.9 und 1.1 (ESM-2.2). Insgesamt weisen diese Ergebnisse darauf hin, dass das Raschmodell für die Items des Mathematiktests in der 6. Klasse eine zumindest akzeptable Modellanpassungsgüte aufweist (Bond & Fox, 2007). Separat nach Schulform ergab sich für SMBG/GY eine WLE-Reliabilität = 0.78/0.73.

Bildungsstandardbasierter Test in der 8. Klasse

In der 8. Klasse enthielt der VERA-Mathematiktest 46 Items zu allen fünf Leitideen. Die vorgesehene Bearbeitungszeit betrug 80 Minuten. Der Test wurde vom Institut zur Qualitätsentwicklung im Bildungswesen konzipiert und auf Grundlage großer bundeslandübergreifender Stichproben normiert. Für die Skalierung wurde ein eindimensionales Rasch-Modell für dichotome Daten genutzt, bei dem die normierten Itemparameter für alle Items herangezogen wurden. Obwohl der Test vorab psychometrisch evaluiert wurde, zeigte sich jedoch, dass für viele Items das Raschmodell nur eine unzureichende Modellpassungsgüte aufwies: So lagen die Outfitmaße für insgesamt 16 der 46 Items ($0.4 \leq 3$ Items ≤ 0.9; $1.3 \leq 12$ Items ≤ 1.8; 1 Item = 3.96) sowie die Infitmaße für insgesamt 13 Items ($0.5 \leq 2$ Items ≤ 0.9; 10 Items = 1.3; 1 Item = 2.10) nicht im modellkonformen Bereich zwischen 0.8 und 1.2 (ESM-2.3; Bond & Fox, 2007). Dennoch wurden die so skalierten Werte für die weiteren Analysen genutzt, um (a) die Generalisierbarkeit auf bundesweit eingesetzte VERA-Mathematiktests nicht einzuschränken und, (b) weil separat nach Schulform zumindest akzeptable Reliabilitäten für die Testleistungen resultierten: WLE-Reliabilität an SMBG/GY = 0.87/0.76.

Tabelle ESM-2.2: Infit- und Outfitmaße für alle Items des VERA-Mathematiktests in der 6. Klasse

Item	Infit	Outfit
[1]	0.9655261	0.9139089
[2]	0.9578081	0.9141151
[3]	0.9993463	0.9880799
[4]	0.9911781	0.9990680
[5]	1.0731535	1.1005325
[6]	0.9533942	0.9140121
[7]	1.0401997	1.2675281
[8]	1.1227308	1.2477335
[9]	1.0570038	1.0977746
[10]	0.9689665	0.9570241
[11]	0.8956471	0.8531192
[12]	0.9087686	0.8813120
[13]	0.9055831	0.8526163
[14]	0.9234048	0.8441616
[15]	1.0863955	1.2545210
[16]	0.9540570	0.9317833
[17]	0.9997612	0.9727582
[18]	1.0096851	1.0238749
[19]	1.1264222	1.1803874
[20]	0.9625314	0.9447389
[21]	0.9207061	0.8208471
[22]	1.0377597	1.0749380
[23]	0.9600174	0.9660576
[24]	0.9652358	0.9669448
[25]	1.0070299	1.0073706
[26]	1.0506547	1.0848627
[27]	1.1478277	1.2728319

Tabelle ESM-2.3: Infit- und Outfitmaße für alle Items des VERA-Mathematiktests in der 8. Klasse

Item	Infit	Outfit	Item	Infit	Outfit
[1]	0.8107473	0.9028382	[28]	0.8084809	0.7336040
[2]	1.0166923	1.0420780	[29]	1.1410577	1.1987852
[3]	1.2118243	1.2013479	[30]	1.3422501	1.5132411
[4]	1.1091271	1.0923626	[31]	1.1143617	1.2656711
[5]	0.9631558	0.9660796	[32]	1.0762577	1.3582171
[6]	1.4825317	1.6630104	[33]	0.8983972	0.8551734
[7]	0.9025407	0.8346811	[34]	0.9409495	0.8786824
[8]	0.9453769	0.9282756	[35]	0.4708875	0.4540165
[9]	1.0052332	0.9976535	[36]	0.6638270	0.5545648
[10]	1.1039616	1.1380912	[37]	1.2555951	1.5513449
[11]	1.1455204	1.0906994	[38]	1.1542376	1.2310857
[12]	2.0944799	3.9589477	[39]	1.2947945	1.4183372
[13]	1.3061293	1.5008230	[40]	0.9530926	0.9515180
[14]	1.2774029	1.3838875	[41]	0.9604791	0.9753019
[15]	1.0078508	1.0353552	[42]	0.9468318	0.9271187
[16]	1.1005747	1.1471362	[42]	1.0707717	1.0971079
[17]	1.0634944	1.1719447	[44]	1.0439071	1.0523013
[18]	1.1174919	1.1836597	[45]	1.2475550	1.3748021
[19]	1.3342107	1.8465711	[46]	1.0204740	1.0216438
[20]	0.9939896	1.0003225	[47]	0.8084809	0.7336040
[21]	1.3430405	1.5192578			
[22]	0.9625654	0.9215638			
[23]	0.9468820	1.0284563			
[24]	1.1986747	1.3679813			
[25]	0.8981067	0.8566297			
[26]	0.9966373	0.9396271			
[27]	0.9442495	0.9061617			

Messinstrumente: Deskriptive Kennwerte zu allen Analysevariablen

Tabelle ESM-2.4: Korrelationsmatrix für die Analysestichprobe auf Basis der imputierten Datensätze mit deskriptiven Kennwerten

		[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1]	T 6	–	.63	.55	.45	*-.02*	-.16	*-.04*
[2]	T 8	.72	–	.55	.54	-.05	-.05	*-.02*
[3]	N 6	.55	.57	–	.59	-.06	*-.03*	*-.02*
[4]	N 8	.45	.55	.57	–	-.08	.09	*-.02*
[5]	Alter 6	-.17	-.22	-.20	-.19	–	-.04	*.01*
[6]	Mäd 6	-.12	-.05	*.01*	.07	-.09	–	*-.02*
[7]	D-MSp 6	-.09	-.08	-.05	-.04	.06	-.03	–
	M_{GY}	0.66	0.70	0.27	0.11	12.48	0.48	0.02
	M_{SMBG}	-0.47	-0.50	-0.19	-0.08	12.67	0.47	0.02
	SD_{GY}	0.82	0.79	0.91	0.98	0.52	-	-
	SD_{SMBG}	0.84	0.82	1.02	1.00	0.66	-	-

Anmerkungen. Korrelationswerte für SuS an GY oberhalb der Hauptdiagonalen. Werte für SuS an SMBG unterhalb der Hauptdiagonalen. In der 1. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung angegeben (bspw. T 6 = Erhebung in der 6. Klasse). T = Mathematiktestleistung, N = Mathematiknote, Alter = Alter in Jahren, Mäd = Dummyvariable für Geschlecht (0 = Jungen; 1 = Mädchen), D-MSp = Dummyvariable für Deutsch als Muttersprache (0 = Deutsch; 1 = andere Sprache als Deutsch). Alle Korrelationen (außer kursiv gesetzte Werte) sind signifikant $p < .05$ (zweiseitig). M = Mittelwert, SD = Standardabweichung. GY = SuS an Gymnasien. SMBG = SuS an Schulen mit mehreren Bildungsgängen (SMBG).

Tabelle ESM-2.5: Korrelationsmatrix für die Moderatoren der Analysestichprobe auf Basis der imputierten Datensätze mit deskriptiven Kennwerten

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
[1]	Rel 6	–	.53	.73	.34	-.26	-.27	.26	*.20*	.28	*-.09*	*-.10*	*-.03*
[2]	Rel 8	.45	–	*.10*	.45	-.70	-.70	.64	.54	.38	*.18*	*.16*	*.12*
[3]	LH 6	.84	.42	–	.39	.35	.21	*.06*	*-.01*	*.16*	*-.10*	*-.20*	*-.07*
[4]	LH 8	.48	.88	.49	–	.00	.24	.73	.59	.30	*.14*	*.02*	*.08*
[5]	LN 6	*.16*	*-.07*	.15	*.11*	–	.77	-.32	-.34	*-.19*	*.03*	*-.13*	*-.04*
[6]	LN 8	*.19*	*-.08*	*.11*	*.13*	.83	–	*-.14*	*-.17*	*-.18*	*-.01*	*-.12*	*-.02*
[7]	βT	*.11*	.65	*.04*	.72	*.03*	*.10*	–	.87	.58	.37	.23	*.18*
[8]	β_{adj}T	*-.02*	.35	*-.01*	.38	*-.01*	*.01*	.71	–	.74	.25	.31	.25
[9]	srT	*.04*	*.03*	*.04*	*.01*	*.00*	*.01*	.29	.64	–	*.03*	*.13*	*.20*
[10]	βN	*-.01*	.26	*.02*	.30	*-.05*	*-.03*	.48	.32	*.12*	–	.73	.58
[11]	β_{adj}N	*-.02*	.13	*.01*	.17	*.04*	*-.01*	.38	.42	.34	.72	–	.91
[12]	srN	*.04*	.13	*.04*	.17	*.06*	*-.01*	.33	.41	.44	.55	.88	–
	M_{GY}	-0.09	-0.85	0.15	0.21	1.42	1.39	0.60	0.43	0.34	0.57	0.22	0.14
	M_{SMBG}	0.03	0.30	-0.05	-0.07	-0.50	-0.49	0.66	0.49	0.40	0.59	0.25	0.18
	SD_{GY}	0.67	0.88	0.73	0.73	0.39	0.41	0.13	0.14	0.10	0.17	0.18	0.12
	SD_{SMBG}	1.09	0.86	1.07	1.07	0.59	0.41	0.13	0.16	0.16	0.17	0.18	0.12

Anmerkungen. Korrelationswerte für SuS an GY oberhalb der Hauptdiagonalen. Werte für SuS an SMBG unterhalb der Hauptdiagonalen. In der 1. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung (bspw. Rel 6 = Erhebung in der 6. Klasse) oder das prognostizierte Kriterium für den Kennwert angegeben (T= Testleistung; N = Note). LN = mittleres Leistungsniveau an der Schule, LH = mittlere Leistungsheterogenität an der Schule, Rel = WLE-Reliabilität, β = Regressionskoeffizient, β_{adj} = adjustierter Regressionskoeffizient, *sr* = Semipartialkorrelationskoeffizient. Alle Korrelationen (außer kursiv gesetzte Werte) sind signifikant $p < .05$ (zweiseitig). *M* = Mittelwert, *SD* = Standardabweichung. $_{GY}$ = SuS an Gymnasien. $_{SMBG}$ = SuS an Schulen mit mehreren Bildungsgängen (SMBG).

ESM-2: Literatur

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates Publishers.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge, Taylor & Francis.

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H. & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments: Psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology*, *7*, 1–14. https://doi.org/10.3389/fpsyg.2016.00154

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Edition.). Chicago, IL: University of Chicago Press.

Robitzsch, A., Kiefer, T. & Wu, M. (2018). *TAM: Test analysis modules*. Verfügbar unter: https://CRAN.R-project.org/package=TAM

Südkamp, A., Pohl, S. & Weinert, S. (2015). Competence assessment of students with special educational needs: Identification of appropriate testing accommodations. *Frontline Learning Reseach*, *3* (2), 1–26. https://doi.org/10.14786/flr.v3i2.130

ESM-3:

Veranschaulichung der Kennwerteverteilungen für die Einzelschulen auf Basis der CR

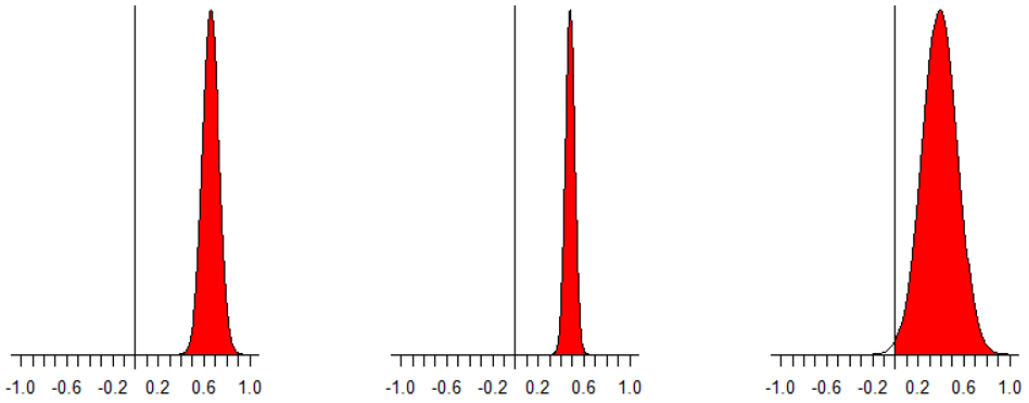
A) SMBG

Kriterium Test:

Forschungsfrage 1): β
100 % der Schulen β > 0

Forschungsfrage 2): β_{adi}
100 % der Schulen $\beta_{adj} > 0$

Forschungsfrage 2): sr
81% der Schulen sr > 0



Kriterium Note:

Forschungsfrage 1): β
93 % der Schulen β > 0

Forschungsfrage 2): β_{adi}
74 % der Schulen $\beta_{adj} > 0$

Forschungsfrage 2): sr
60 % der Schulen sr > 0

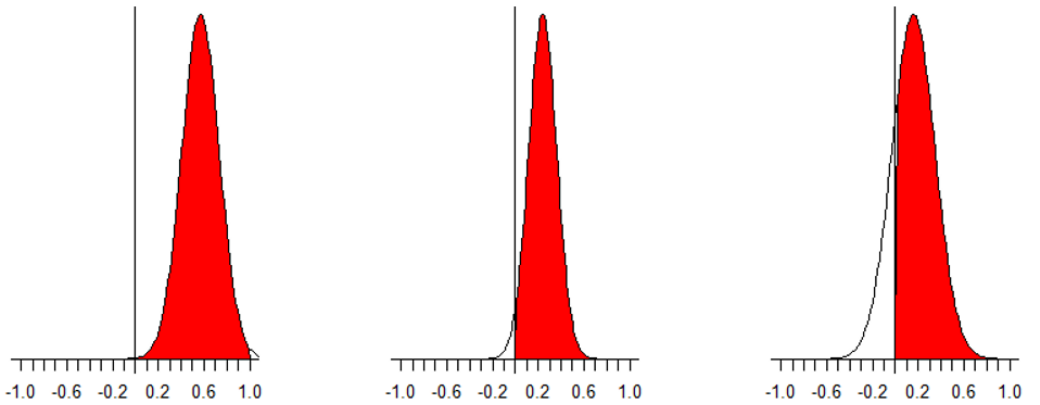
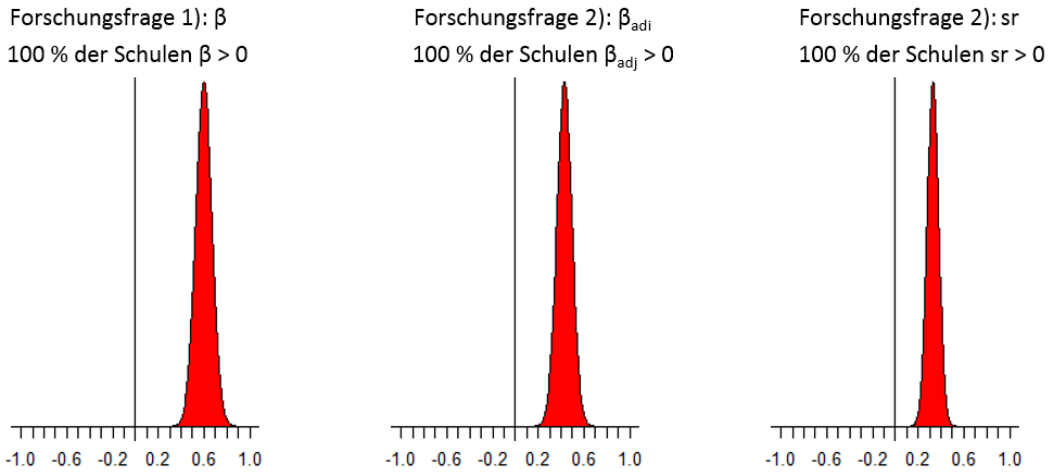


Abbildung ESM-3.1. Darstellung der Kennwerteverteilung für die Einzelschulen der SMBG auf der Grundlage der berechneten 95 %-CR für die Kennwerte zur Prognosegüte. Pro Kennwert wird angegeben, wie hoch der prozentuale Anteil der Einzelschulen war, deren Kennwert für die Prognosegüte größer war als 0.

B) GY

Kriterium Test:



Kriterium Note:

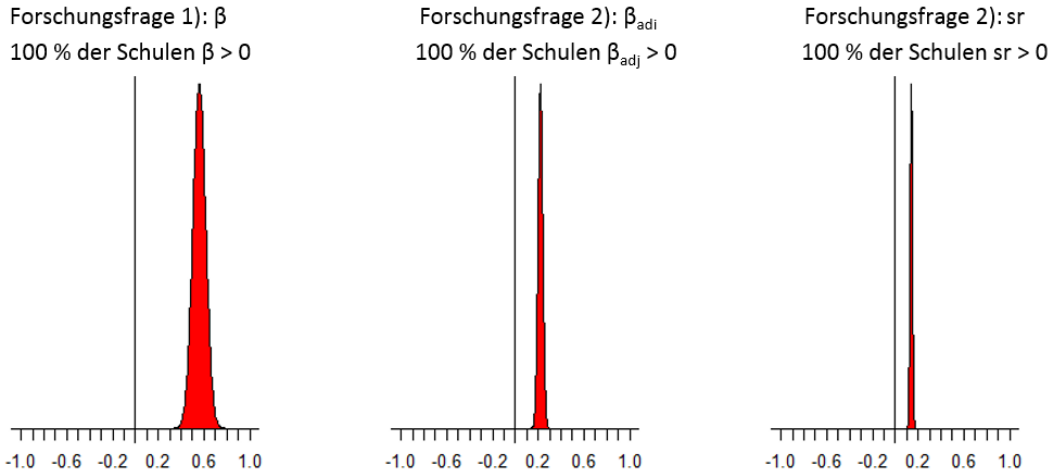


Abbildung ESM-3.2. Darstellung der Kennwerteverteilung für die Einzelschulen der GY auf der Grundlage der berechneten 95 %-CR für die Kennwerte zur Prognosegüte. Pro Kennwert wird angegeben, wie hoch der prozentuale Anteil der Einzelschulen war, deren Kennwert für die Prognosegüte größer war als 0.

ESM-4:

Ergänzende Erläuterungen für den Zusammenhang zwischen Reliabilität und Leistungsniveau auf Schüler- und Schulebene

GY

Aus der Abbildung ESM-4.1B wird deutlich, dass die VERA-Mathematiktests an GY vor allem für leistungsstarke SuS weniger reliabel sind. Infolgedessen fallen die schulspezifischen Reliabilitäten an GY mit leistungsstarker Schülerschaft tendenziell niedriger aus (Abbildung ESM-4.2B). Dieser lineare Zusammenhang zwischen der schulspezifischen Reliabilität und dem schulspezifischen Leistungsniveau in der 6./8. Klasse beträgt r = -.26/-.27 (ESM-2.5).

SMBG

Aus der Abbildung ESM-4.1A wird deutlich, dass die VERA-Mathematiktests an SMBG einerseits für deutlich leistungsschwache und andererseits für leistungsstarke SuS weniger reliabel sind. Infolgedessen fallen die schulspezifischen Reliabilitäten an SMBG mit deutlich leistungsschwacher, sowie leistungsstarker Schülerschaft tendenziell niedriger aus (Abbildung ESM-4.2A). Dieser tendenziell nicht lineare Zusammenhang kann durch den Korrelationskoeffizienten nicht adäquat abgebildet werden, weshalb möglicherweise niedrige Koeffizienten an SMBG zwischen der schulspezifischen Reliabilität und dem schulspezifischen Leistungsniveau in der 6./8. Klasse von r = .16/-.08 resultierten (ESM-2.5).

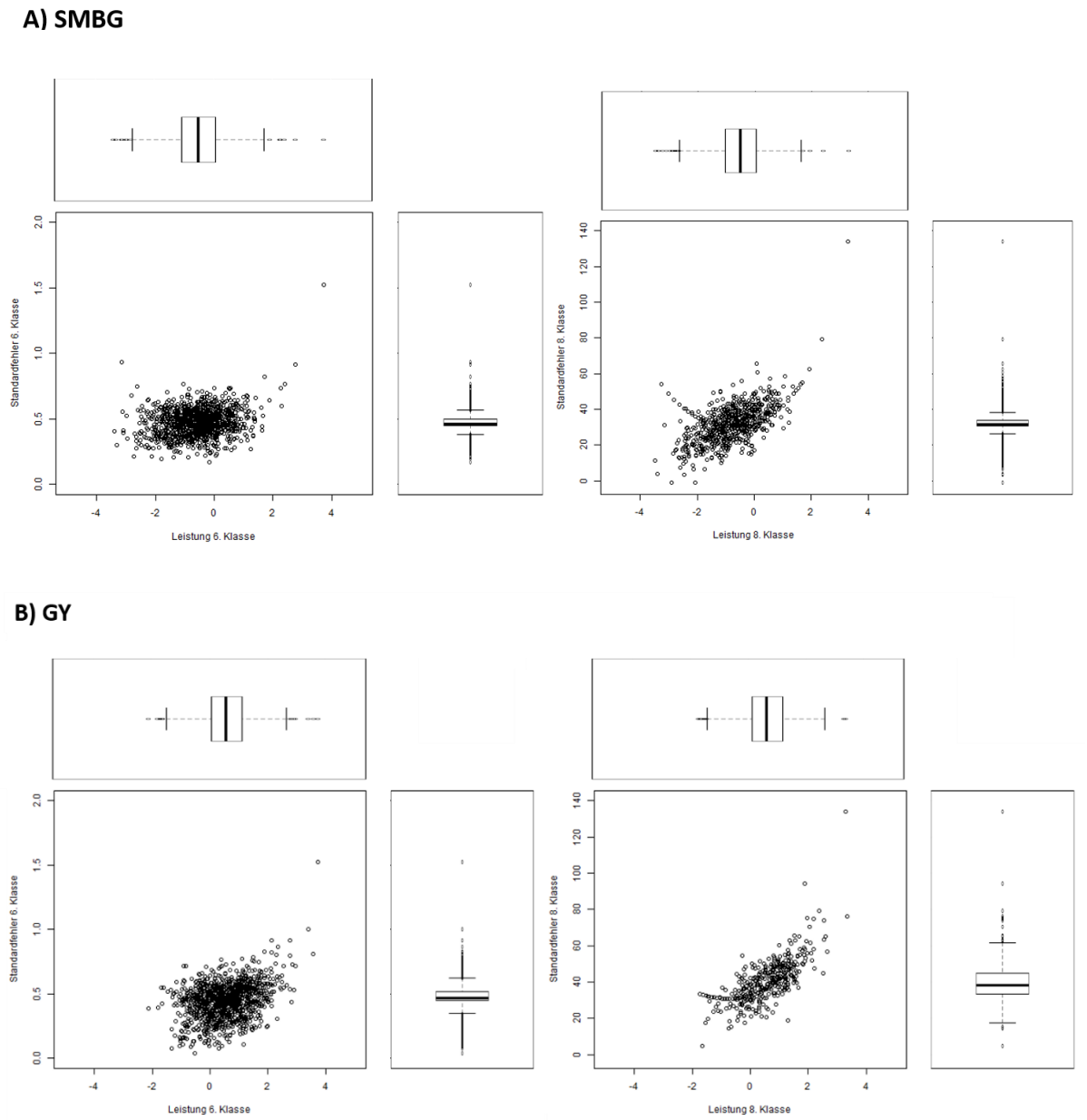
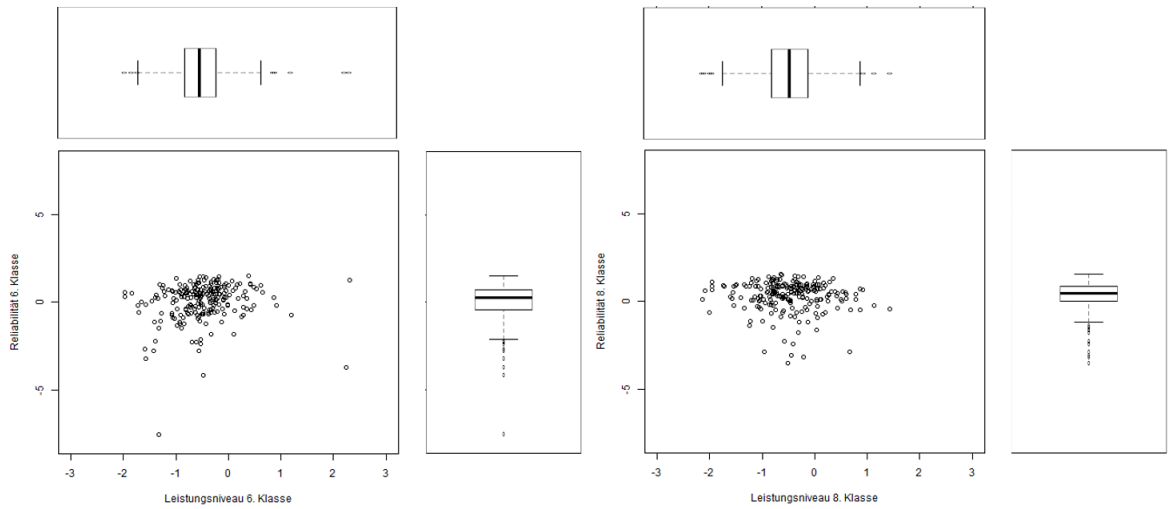


Abbildung ESM-4.1. Schülerebene: Streudiagramme zur Veranschaulichung des Zusammenhangs zwischen dem Leistungsniveau der SuS und dem jeweiligen Standardfehler für die Messzeitpunkte in der 6. und 8. Klasse separat für die Gesamtstichprobe an (A) SMBG bzw. (B) GY. Exemplarische Darstellung für die Daten des ersten imputierten Datensatzes der insgesamt 15 imputierten Datensätze). Zudem Abbildung der jeweiligen Box-Plots für die dargestellten Merkmale.

A) SMBG



B) GY

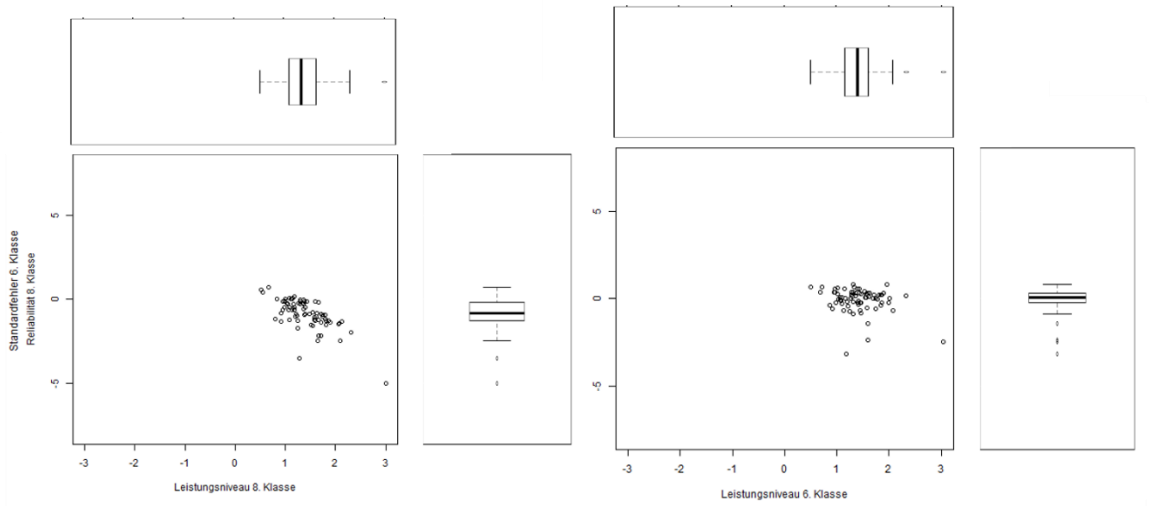


Abbildung ESM-4.2. Schulebene: Streudiagramme zur Veranschaulichung des Zusammenhangs zwischen dem schulspezifischen Leistungsniveau und der schulspezifischen Reliabilität für die Messzeitpunkte in der 6. und 8. Klasse separat für die Gesamtstichprobe an (A) SMBG bzw. (B) GY. Zudem Abbildung der jeweiligen Box-Plots für die dargestellten Merkmale.

ESM-5: Funnel plots für die jeweiligen Kennwerte der Prognosegüte

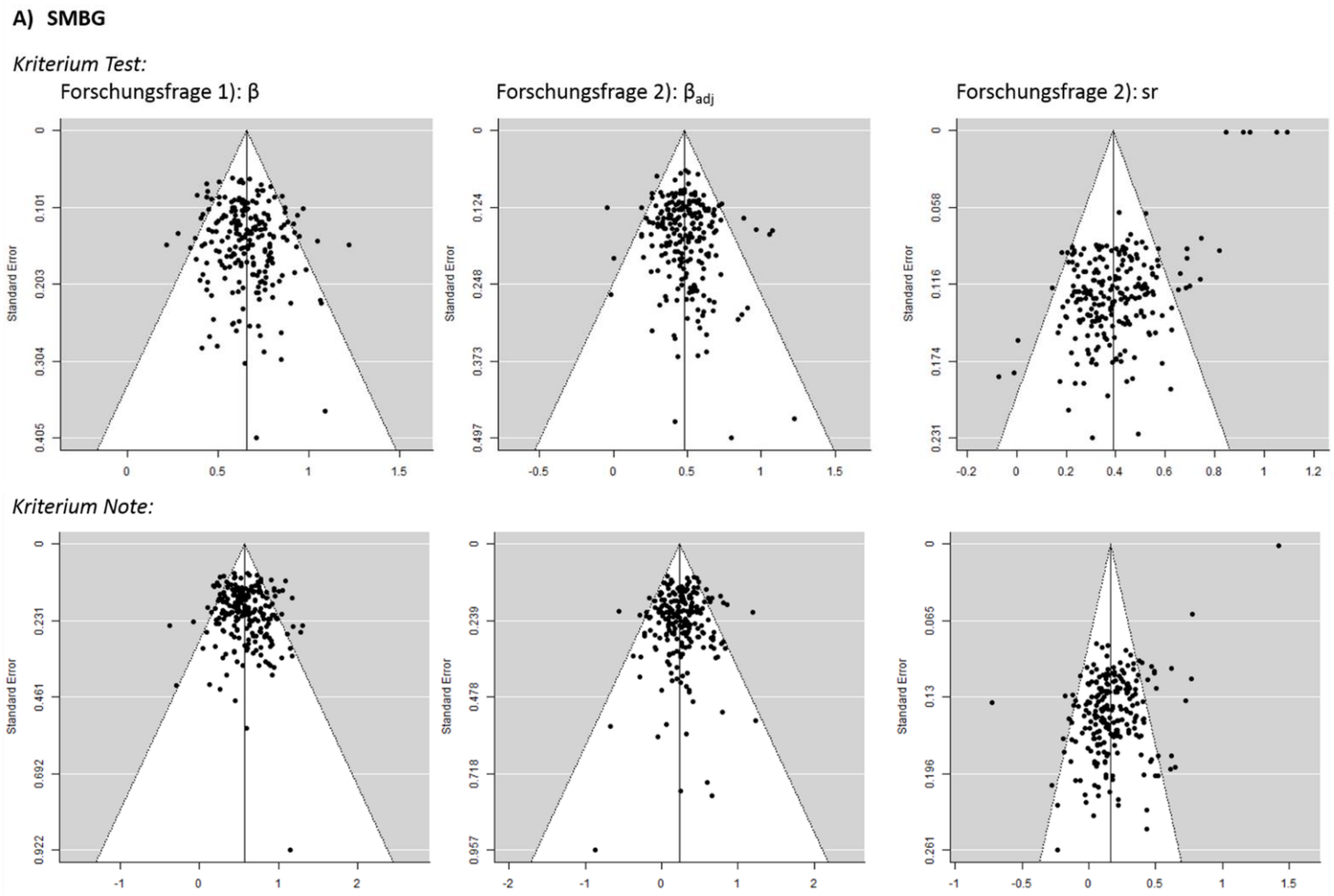


Abbildung ESM-5.1. Funnel plots zur Veranschaulichung der Kennwerteverteilung für die jeweiligen Einzelschulen in Abhängigkeit des jeweiligen Standardfehlers für SMBG. Bei der Berechnung des sr für SMBG resultierten für 5 Schulen für die Prognose der Testleistung und für eine Schule für die Prognose von Schulnoten negative Varianzen. Diese setzten wir auf einen kleinen positiven Wert (= 0.000001; s. Schmidt & Hunter, 2015), um sie in die Analysen einbeziehen zu können. Dies führt jedoch zu einer leichten systematischen Erhöhung der Heterogenität.

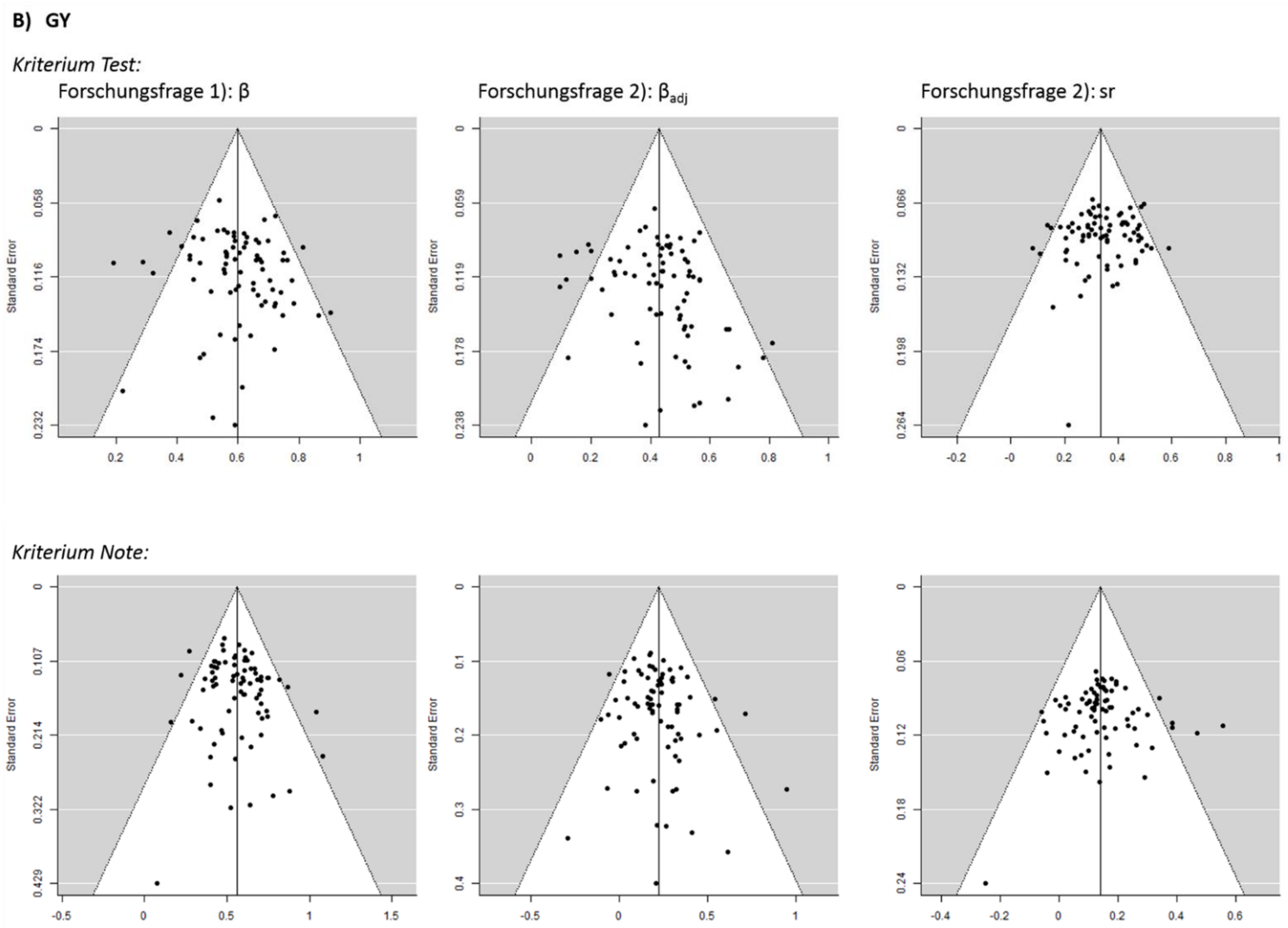


Abbildung ESM-5.2. Funnel plots zur Veranschaulichung der Kennwerteverteilung für die jeweiligen Einzelschulen in Abhängigkeit des jeweiligen Standardfehlers für GY.

ESM-5: Literatur

Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis: correcting error and bias in research findings* (3rd ed.). Thousand Oaks, California: SAGE.

ESM-6: Methodenvergleich zur Ermittlung der Prognosegüte

Modell		(1) Gesamt-stichprobe	(2) Mehrebenenanalyse		(3) Metaanalytischer Ansatz	
A) SMBG						
Kriterium: Test		Mittelwert	Mittelwert	τ	Mittelwert	τ
	β	.70	.67	.07	.66	.07
	β_{adj}[a]	.56	.49	.04	.48	.04
	sr[a]	.48	-	-	.39	.15
Kriterium: Note						
	β	.54	.58	.13	.57	.16
	β_{adj}[a]	.23	.24	.08	.24	.12
	sr[a]	.16	-	-	.16	.18
B) GY						
Kriterium: Test		Mittelwert	Mittelwert	τ	Mittelwert	τ
	β	.61	.60	.07	.60	.07
	β_{adj}[a]	.45	.43	.06	.43	.07
	sr[a]	.39	-	-	.33	.05
Kriterium: Note						
	β	.54	.56	.05	.56	.05
	β_{adj}[a]	.21	.21	.03	.22	.00[b]
	sr[a]	.14	-	-	.14	< .01

Anmerkungen. Methodenvergleich von (1) bisher verbreitetem Ansatz mit der Ermittlung der Kennwerte auf Grundlage der Gesamtstichprobe ohne Berücksichtigung der Heterogenität zwischen Einzelschulen, (2) Berücksichtigung der Heterogenität zwischen Einzelschulen über die Anwendung von Mehrebenenanalysen sowie (3) Berücksichtigung der Heterogenität zwischen Einzelschulen über die Anwendung des metaanalytischen Ansatzes, welcher im Rahmen dieser Studie fokussiert wurde. Metrik der Prädiktor- und Kriteriumvariable (s. Spezifikation im Text). SMBG = Schulen mit mehreren Bildungsgängen, τ = Standardabweichung des jeweiligen Kennwerts, β = Regressionskoeffizient, β_{adj} = adjustierter Regressionskoeffizient, sr = Semipartialkorrelationskoeffizient.

[a] Kontrolle für die Halbjahresnote in Mathematik in der 6. Klasse. [b] τ = 0 trotz Restricted Maximum Likelihood Verfahren (REML; Veroniki et al., 2016).

Literatur

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G. et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7 (1), 55–79. https://doi.org/10.1002/jrsm.1164

Anhang C: Manuskript und elektronische Supplemente - Studie III

STUDIE III: Wie gut lassen sich mit bildungsstandardbasierten Kompetenztests Kinder identifizieren, die wichtige Bildungsergebnisse im Verlauf der Schulkarriere verfehlen? Ergebnisse zweier Längsschnittstudien zur Klassifikationsgüte in den Fächern Mathematik und Deutsch.

Die Teilstudie ist als Zeitschriftenbeitrag in Begutachtung:

Fuchs, G., Nachtigall, C., Harych, P. & Brunner, M. *(eingereichte Manuskriptfassung in Begutachtung)*. Wie gut lassen sich mit bildungsstandardbasierten Kompetenztests Kinder identifizieren, die wichtige Bildungsergebnisse im Verlauf der Schulkarriere verfehlen? Ergebnisse zweier Längsschnittstudien zur Klassifikationsgüte in den Fächern Mathematik und Deutsch. *Diagnostica*.

(Diese Artikelfassung wurde in dieser Form noch nicht für eine Veröffentlichung angenommen.)

Zusammenfassung

In zwei Längsschnittstudien (N_1 = 10 939, N_2 = 430) untersuchten wir, inwiefern bildungsstandardbasierte Mathematik- und Deutschtests in der 3. bzw. 4. Klasse geeignet sind, Kinder zu identifizieren, die „gefährdet“ sind, zukünftige Bildungsergebnisse (bis zu 5 Jahre später) zu verfehlen: Verfehlen der (a) Mindeststandards bzw. Regelstandards und (b) Schulnoten 1 bis 3. Als Vergleichsmaßstab zur Bewertung der Klassifikationsgüte (Sensitivität, positiv prädiktiver Wert) nutzten wir Schulnoten. Zudem prüften wir, inwiefern die Kombination von Testergebnis und Note die Güte verbessert. Die Ergebnisse zeigten, dass die Sensitivität der Tests in beiden Fächern auf Basis beider Schwellenwerte zum Teil (deutlich) höher ausfällt im Vergleich zur Note. Die Ergebnisse in Abhängigkeit vom Fach und der Schulform jedoch deutlich variierten. Kombinierte man Testergebnis und Note, konnten diagnostische Entscheidungen zum Teil verbessert werden. Zusammenfassend weisen die Befunde daraufhin, dass im Rahmen der bundesweiten Vergleichsarbeiten Lehrkräfte Informationen erhalten, die die Identifizierung von Förderbedarf verbessern können.

Schlüsselwörter: Bildungsstandards; Sensitivität; positiv prädiktiver Wert; Screening; Diagnostik

Abstract

Drawing on data of two longitudinal studies (main study: N_1 = 10 939; additional study: N_2 = 430) we examined in different federal states of Germany the diagnostic accuracy of proficiency tests in mathematics and German reading comprehension that assess national educational standards. Specifically, we investigated whether proficiency levels assessed in the 3rd or 4th grade can identify students at risk of failing important educational outcomes up to 5 years later: (a) minimum proficiency levels or average proficiency levels and (b) grades 1, 2 and 3. We analyzed the diagnostic accuracy (in terms of sensitivity and positive predictive value) of proficiency levels separately as well as in in combination with grades. Our results showed that the sensitivity obtained for students' proficiency levels was often (considerably) higher than those obtained for grades. However, the achieved level of diagnostic accuracy depended mainly on the subject and school type. In many cases, diagnostic accuracy was improved by a combination of proficiency levels and grades. To conclude, state-wide assessment programs offer teachers information on students' proficiency levels to improve their diagnostic decisions on which students are in need of additional learning support.

Keywords: standard of education; sensitivity; positive predictive value; screening; diagnostic

Die kriteriumsorientierte Leistungsmessung ist seit den 1970er-Jahren ein zentrales Forschungsgebiet der pädagogisch-psychologischen Diagnostik (Ingenkamp & Lissmann, 2008; Klauer, 1987). Mit den Bildungsstandards (Kultusministerkonferenz [KMK], 2006) liegt nun seit einigen Jahren eine verbindliche, bundesweit einheitliche kriteriale Bezugsnorm vor. Die Bildungsstandards formulieren fachbezogene Kompetenzerwartungen, die Schülerinnen und Schüler (SuS) bis zu einem bestimmten Zeitpunkt ihrer Bildungskarriere erreicht haben sollen. Informationen darüber, ob die eigenen SuS diese Erwartungen erfüllen, können Lehrkräfte mithilfe von bildungsstandardbasierten Kompetenztests erhalten, zum Beispiel den sogenannten Vergleichsarbeiten (VERA). Lehrkräfte der 3. und 8. Jahrgangsstufe an allen öffentlichen Schulen in Deutschland sind verpflichtet, jährlich in mindestens einem Fach VERA durchzuführen (KMK, 2012). Im Rahmen einer Befragung des Ländervergleichs im Jahr 2011 stimmten 45 % der Grundschullehrkräfte zu, dass VERA-3-Tests „eine gute Grundlage für die Planung individueller Fördermaßnahmen bieten“ (Richter & Böhme, 2014). Aus wissenschaftlicher Sicht wird das Potenzial dieser Tests für die individualdiagnostische Anwendung vor allem darin gesehen, dass diese im Sinne eines „Screenings“ Hinweise für eine weiterführende Diagnostik geben können (u. a. Köller & Reiss, 2013). Um die Qualität dieser Screeningfunktion empirisch zu beurteilen, ist es wichtig zu wissen, inwiefern Kinder identifiziert werden können, die „gefährdet“ sind, wichtige Bildungsergebnisse auf ihrem weiteren Bildungsweg zu verfehlen. Bisher fehlt es jedoch an empirischen Studien, die untersuchen, wie gut diese Identifikation gelingt. Dieser Forschungsfrage gehen wir im Rahmen zweier Längsschnittstudien nach, indem wir die prognostische Klassifikationsgüte von bildungsstandardbasierten Tests über mehrere Jahre bestimmen.

Forschungsstand zur prognostischen Klassifikationsgüte standardisierter Schulleistungstests und von Schulnoten

Fachspezifische Kompetenzen bzw. das Vorwissen zu einem früheren Zeitpunkt gehören zu den besten Prädiktoren für den zukünftigen schulischen Kompetenz- und Wissenserwerb (Helmke & Weinert, 1997). So sagten bildungsstandardbasierte Tests (Bista-Tests) schulische Erfolgskriterien innerhalb der Primarstufe (Fuchs & Brunner, 2017) und Sekundarstufe (Graf, Harych, Wendt, Emmrich & Brunner, 2016) substanziell vorher und zwar auch dann, wenn für weitere leistungsprädiktive Merkmale (z. B. Schulnoten) kontrolliert wurde. Diese Befunde basieren auf Regressionsanalysen, die Zusammenhänge zwischen Prädiktoren und

Kriterien auf Grundlage des gesamten kontinuierlichen Merkmalsspektrums quantifizieren. Im Rahmen der Individualdiagnostik (z. B. zur Identifikation von Förderbedarf) ist es grundsätzlich jedoch erforderlich zu entscheiden, ob eine Person einen bestimmten Wert (= Schwelle) überschritten hat oder nicht, bzw. ein Kriterium erreicht hat oder nicht. Die Individualdiagnostik ist damit eng an die Klassifikation von Personen gebunden. Die Klassifikationsgüte von Tests kann hierzu mit zahlreichen Kennwerten, wie zum Beispiel der Sensitivität (s. a. den Abschnitt „Statistische Analysemodelle“) bewertet werden (u. a. Marx, 1992; Gersten et al., 2012).

Obwohl die Identifikation von Kindern mit zusätzlichem schulischen Förderbedarf sicherlich eine wichtige diagnostische Zielstellung von Leistungstests ist, gibt es kaum belastbare Befunde zur Klassifikationsgüte standardisierter Schulleistungstests im deutschsprachigen Raum. Der Forschungsstand [Elektronisches Supplement (ESM)-1] in Mathematik und Deutsch (i. w. S. für das Lesen) zur Prognose von Testleistungen und Noten lässt sich wie folgt zusammenfassen: (a) In den meisten Studien erfolgte die Klassifikation der Schülerleistungen in der Primarstufe. Für Bista-Tests gibt es bislang keine Studien. (b) Für die Sekundarstufe liegen ebenso keine Studien zu Bista-Tests vor. Die Ergebnisse von Graf und Kollegen (2016) erlauben es, die Klassifikationsgüte von VERA-8-Tests hinsichtlich der 2 Jahre später erreichten Noten in den zentralen Prüfungen zum Mittleren Schulabschluss abzuschätzen: Die von uns berechneten Werte für die Sensitivität (*Sen*) liegen zwischen 5 % und 43 % (ESM-1, S. 4 und 6, zur Interpretation s. den Abschnitt „Statistische Analysemodelle“). (c) Die Prognosedauer bisheriger Studien betrug maximal 3 Jahre. (d) Die Schwellen, die zur Klassifikation genutzt werden, basieren ausnahmslos auf sozialen Bezugsnormen: So werden Kinder als „gefährdet“ eingestuft, wenn sie im Vergleich zu den restlichen Kindern die schwächsten Kompetenzen erreichen. Meist wird hier der Prozentrang 15 als Schwelle herangezogen, also wenn lediglich 15 % aller getesteten Kinder schwächere oder gleich schwache Leistungen aufzeigen. (e) In nahezu allen Studien zur prognostischen Klassifikationsgüte von Noten, wurde unterschieden, ob SuS die Noten 1 bis 3 erreichten bzw. verfehlten. (f) Nur für 2 von 19 Tests wurde geprüft, inwiefern sich die Klassifikationsgüte durch Kombination mehrerer Leistungsprädiktoren verbessert. (g) Für die Lesekompetenz in Deutsch liegen deutlich weniger Befunde vor als für die Mathematikkompetenz.

Mit Blick auf die Schulnoten zeigten zahlreiche Studien auf Basis von Korrelations- und Regressionsanalysen, dass Noten eine substantielle Vorhersage von späteren Schulleistungen

ermöglichen (für Überblick s. Lintorf, 2012). Die Frage zur Klassifikationsgüte von Noten wurde unseres Wissens im deutschsprachigen Raum bislang nur in der Querschnittsstudie von Hoffmann und Böhme (2017) mit Kindern in der 4. Klassenstufe für das Fach Deutsch untersucht: „Schwache Noten“ (die Noten 4, 5 und 6) wiesen eine *Sen* von 32 % auf, Kinder zu identifizieren, die den Mindest- bzw. Regelstandards im Zuhören und Lesen verfehlten.

Schwellen(-werte), Kriterien und Strategien

Das spätere Erreichen bzw. Verfehlen von Bildungsergebnissen bezeichnen wir (dem üblichen Jargon im Rahmen der klassifikatorischen Diagnostik folgend) als das Erreichen bzw. Verfehlen von Kriterien. Prädiktoren und Kriterien werden jeweils in dichotomer Form betrachtet: Schwelle beim Prädiktor nicht überschritten (Kind „gefährdet“) vs. Schwelle überschritten (Kind „nicht gefährdet“); Kriterium verfehlt (Bildungsergebnis verfehlt) vs. Kriterium nicht verfehlt (Bildungsergebnis nicht verfehlt). Den Empfehlungen der Teststandards folgend (American Educational Research Association [AERA], American Psychological Association & National Council on Measurement in Education, 2014) stellen wir im Folgenden detailliert die Rationale dar, wie wir für Prädiktoren und Kriterien die Schwellen begründen. Diese Erläuterungen sind auch deshalb wichtig, da die Indizes zur Bewertung der Klassifikationsgüte von den Selektionsraten bezüglich der Prädiktoren bzw. den Basisraten der Kriterien und damit von den verwendeten Schwellen abhängig sind (Bossuyt et al., 2015).

Bildungsstandardbasierte Tests als Prädiktor

Wichtige Schwellen stellen die erreichten fachspezifischen (Kompetenz-)Stufen dar, die mit bildungsstandardbasierten Mathematik- und Deutschtests für die Primarstufe erfasst werden können. Diese Stufen beschreiben die Kompetenzerwartungen an Kinder in der 4. Klasse. Das Erreichen bzw. Verfehlen des Mindeststandards ist eine zentrale Schwelle. Kinder, die den Mindeststandard in Mathematik verfehlen, klassifizieren wir als „gefährdet“, denn sie haben „die eigentlichen Ziele des Mathematikunterrichts in der Grundschule . . . weitestgehend noch nicht erreicht“ (KMK, 2013a, S. 12). In Deutsch (Lesen) bleiben Kinder auf dieser Stufe „deutlich hinter den Erwartungen der Bildungsstandards zurück. Es ist davon auszugehen, dass der erfolgreiche Übergang in die Sekundarstufe I nur unter Einsatz intensiver Fördermaßnahmen gelingen wird“ (KMK, 2013b, S. 10). Kinder, die in Mathematik bzw.

Deutsch den Mindeststandard oder höher erreichen, klassifizieren wir als „nicht gefährdet“. Nachfolgend bezeichnen wir die (alleinige) Nutzung der Testergebnisse zur Prognose späterer Bildungsergebnisse als T_M für Mathematik und T_D für Deutsch. Als alternative Schwelle zur Identifizierung „gefährdeter“ Kinder nutzten wir zudem das Verfehlen des Regelstandards. Dieser definiert die fachspezifischen Erwartungen der KMK an die Kompetenzen, über die Kinder im Durchschnitt am Ende der 4. Klasse verfügen sollten (KMK, 2013a). Aufgrund der höher gesetzten Schwelle, erhöht sich die Selektionsrate, die die *Sen* des Testverfahrens im Rahmen eines „Screenings“ begünstigt. Mit der alternativen Schwelle ist es somit möglich abzuschätzen, inwiefern die Klassifikationsgüte von unterschiedlichen, aber dennoch praxisrelevanten Entscheidungskriterien abhängt.

Schulnoten als Prädiktor

Für die Klassifikation der Kinder auf Basis der Mathematik- bzw. Deutschnoten berücksichtigen wir neben inhaltlichen Aspekten auch die Prävalenz bzw. Selektionsrate als statistische Randbedingung. Auf der Notenskala werden die Noten 5 und 6 als „nicht bestanden“ gewertet. Somit würde sich die Note 4 als Schwelle anbieten. Jedoch werden die Noten 5 und 6 in der Primarstufe im Zeugnis nur äußerst selten vergeben; ihre Prävalenz liegt unter 6 % (Bos, Tarelli, Bremerich-Vos & Schwippert, 2012; Hochweber, 2010; Tab. ESM-2 bzw. ESM-5.2). Eine geringe Prävalenz wirkt sich nachteilig auf die *Sen* von Noten aus. Daher haben wir im Einklang mit den meisten bisherigen Studien (ESM-1) die Note 3 als Schwelle gewählt: Kinder, die im Halbjahreszeugnis die Noten 4, 5 oder 6 erhalten haben, klassifizieren wir als „gefährdet“, spätere Bildungsergebnisse zu verfehlen. Kinder, die die Noten 1, 2 oder 3 erhalten haben, klassifizieren wir als „nicht gefährdet“. Nachfolgend bezeichnen wir die (alleinige) Nutzung der Note zur Prognose mit N_M für Mathematik und N_D für Deutsch.

Strategien zur Kombination von Prädiktoren

In der Regel werden diagnostische Entscheidungen auf der Grundlage mehrerer Informationsquellen getroffen (Koretz, 2003). Folglich untersuchen wir zwei potentielle Strategien zur Kombination von Testergebnis und Note: Auf der Grundlage der ODER-Strategie (T_MoN_M in Mathematik bzw. T_DoN_D in Deutsch) werden all diejenigen Kinder als „gefährdet“ klassifiziert, die die Schwelle des Tests *oder* der Note desselben Faches

unterschreiten. Mit dieser Strategie werden bspw. in Mathematik mit der Orientierung am Mindeststandard diejenigen Kinder als „gefährdet“ klassifiziert, die in Mathematik entweder (a) im Test den Mindeststandard verfehlen oder (b) die Note 4, 5 oder 6 erhalten sowie (c) Kinder, auf die beides zutrifft. Hingegen werden bei der UND-Strategie ($T_M u N_M$ bzw. $T_D u N_D$) alleinig diejenigen Kinder als „gefährdet“ klassifiziert, die die Schwelle des Tests *und* der Note desselben Faches unterschreiten, zum Beispiel diejenigen Kinder, die in Mathematik sowohl im Test den Mindeststandard verfehlen und zugleich die Note 4, 5 oder 6 erhalten.

Interessant ist, dass beide Strategien in der pädagogischen Praxis eine Rolle spielen (u. a. Hellrung & Hartig, 2013). So gaben einige Lehrkräfte in der Studie von Maier (2009) an, die Ergebnisse der VERA-Tests als zusätzliche Diagnoseinformation zu einzelnen SuS zu nutzen; diese Angaben stehen also im Einklang mit der Verwendung einer ODER- bzw. einer UND-Strategie. Andere Lehrkräfte gaben an, die Ergebnisse lediglich als Information zu nutzen, die bisheriges Diagnosewissen bestätigt; dies entspricht einer UND-Strategie.

Prognostizierte Kriterien

Die Klassifikationsgüte von Bista-Tests und Noten untersuchen wir im Hinblick auf das Verfehlen von zwei fachbezogenen Bildungsergebnissen: In der Hauptstudie für das Verfehlen (1) des Mindeststandards und (2) der Noten 1 bis 3; in der Ergänzungsstudie für das Verfehlen (1) des Regelstandards und (2) der Noten 1 bis 3. Der Grund für die unterschiedlich gesetzten Standards ist, dass die Kriterien in unterschiedlichen Klassenstufen erhoben wurden: In der Hauptstudie in der 8. Klasse der Sekundarstufe und in der Ergänzungsstudie in der 5. bzw. 6. Klasse der Primarstufe.

Um die Schwelle für die Tests in der 8. Klasse zu definieren, verwenden wir das integrierte Kompetenzstufenmodell für die Sekundarstufe in Mathematik bzw. Deutsch im Bereich Lesen (KMK, 2013c, 2014). Jugendliche, deren Kompetenzen unterhalb des Mindeststandards für den Mittleren Schulabschluss liegen (dies entspricht dem Regelstandard für den Hauptschulabschluss), verfehlen die Erwartungen für die 8. Klasse. Obwohl die Standards erst für das Ende der Sekundarstufe I (10. Klasse) definiert sind, sind diese Schwellen praxisrelevant: So wird z. B. in Mathematik erwartet, dass Kompetenzen unterhalb dieser Schwelle „typischerweise bis etwa zum siebten Schuljahr [also bereits vor der 8. Klasse] des Hauptschulbildungsganges erreicht“ werden (KMK, 2013c, S. 62). Vor allem für leistungsstärkere Schülergruppen, wie bspw. Gymnasiasten, ist a priori zu erwarten, dass nur

sehr wenige Jugendliche dieses Kriterium verfehlen werden. Daher ist davon auszugehen, dass die Klassifikationsgüte am Gymnasium eher gering ausfallen wird. Um für die Ergänzungsstudie die Schwelle für Kinder in der 5. bzw. 6. Klasse zu bestimmen, ziehen wir das Kompetenzstufenmodell für die Primarstufe heran, da die Kinder Schulen in einem Bundesland mit Primarstufe bis zur 6. Klasse besuchten. Dabei definieren wir als Schwelle für die 5. bzw. 6. Klasse jeweils den Regelstandard, da dieser die Leistungserwartung der KMK definiert, die im Durchschnitt von den Kindern bereits bis zum Ende der 4. Klasse in der Grundschule erfüllt werden sollte.

Bezüglich der Noten als Kriterium wurden in beiden Studien dieselben Schwellen definiert: Kinder bzw. Jugendliche, die die Noten 4, 5 oder 6 erhalten haben, haben das Kriterium verfehlt (Prävalenz der Noten Tab. ESM-2 bzw. ESM-5.2).

Forschungsfragen

Bista-Tests werden flächendeckend im Rahmen von VERA im Schulkontext eingesetzt. Sie können genutzt werden, um zum Beispiel im Kontext von „Screenings“ die Identifikation von Förderbedarf bei Kindern zu unterstützen (Köller & Reiss, 2013). Bista-Tests verorten hierzu SuS auf den Kompetenzstufen der Bildungsstandards. Obwohl es im Kontext klassifikatorischer Diagnostik nachdrücklich empfohlen wird (AERA, 2014; Gersten et al., 2012; Marx, 1992), fehlt es bislang jedoch an Studien, die die Klassifikationsgüte von Bista-Tests empirisch untersuchen. Diese Forschungslücke gehen wir mit der vorliegenden Studie an, indem wir systematisch analysieren, wie gut Bista-Tests „gefährdete“ Kinder identifizieren können. Hierzu gehen wir im Rahmen von zwei Längsschnittstudien drei übergreifenden Forschungsfragen nach: (1) Inwiefern sind die Ergebnisse von Bista-Tests in Mathematik (T_M) bzw. Deutsch im Bereich Lesen (T_D) in der 3. bzw. 4. Klasse geeignet, „gefährdete“ Kinder zu identifizieren, die einige Jahre später wichtige Bildungsergebnisse in diesen Fächern (Kompetenz- bzw. Notenstufe) verfehlen? Hierzu vergleichen wir die Sensitivität (*Sen*) und den positiv prädiktiven Wert (*ppW*) für zwei alternative Schwellen zur Identifikation von „gefährdeten“ Kindern: das Verfehlen des (a) Mindeststandards und (b) des Regelstandards. (2) Wie fällt die Klassifikationsgüte von Bista-Tests im Vergleich zu alternativen praxisrelevanten Prädiktoren aus und zwar den Noten in Mathematik (N_M) bzw. Deutsch (N_D)? (3) Inwiefern lässt sich durch die Kombination von Bista-Testergebnis und

Note in Mathematik (T_MuN_M bzw. T_MoN_M) bzw. Deutsch (T_DuN_D bzw. T_DoN_D) die Identifikation „gefährdeter“ Kinder verbessern?

Methode

Stichprobe und Prozeduren

Im Fokus dieses Artikels steht die Hauptstudie zur Klassifikationsgüte von VERA-Tests und Noten in den Fächern Deutsch und Mathematik für den Prognosezeitraum von der 3. bis zur 8. Klasse. Im Rahmen der Ergänzungsstudie untersuchen wir, inwiefern sich die Ergebnisse im Fach Mathematik für Prognosen ab der 4. Klasse auf die 5. bzw. 6. Klasse innerhalb der Primarstufe replizieren lassen (Details zur Ergänzungsstudie mit N = 430 in ESM-5).

Für die Hauptstudie nutzten wir einen Archivdatensatz mit Längsschnittdaten von SuS, die an den Vergleichsarbeiten im selben Bundesland[4] in der Primar- und Sekundarstufe teilnahmen. Der Archivdatensatz wurde rückwirkend von den Daten in der 8. Klasse erstellt. Folglich gibt es keinen systematischen Stichprobenausfall, der üblicherweise in Längsschnittstudien entsteht, in denen der Datensatz ausgehend von der ersten Erhebung prospektiv erstellt wird. Die erste Erhebung fand in der 3. Klasse (VERA 3) im Mai 2011 und die zweite in der 8. Klasse (VERA 8) im Februar 2016 statt. Alle öffentlichen Schulen waren zur Durchführung der Tests in beiden Klassenstufen verpflichtet. Von den Analysen schlossen wir Kinder mit sonderpädagogischem Förderbedarf (Hören, Sehen, körperlich-motorische Entwicklung, Lernen, Sprache und soziale Entwicklung) aus, da die Tests für diese Kinder nur bedingt geeignet sind. Darüber hinaus bezogen wir nur Kinder bzw. Jugendliche in die Analysen ein, für die in der 3. Klasse für alle Prädiktoren (Testergebnis und Note in Mathematik und Deutsch) Daten vorlagen. Es zeigten sich lediglich kleine Unterschiede in den Leistungsmaßen zwischen den Kindern bzw. Jugendlichen der Analysestichprobe und jenen, die einen oder mehr fehlende Werte bei den Prädiktoren aufwiesen (Details s. Tab. ESM-3). Insgesamt umfasste die Analysestichprobe 10 939 SuS (49 % Mädchen; Durchschnittsalter in der 3. Klasse 9.5 Jahre, SD = 0.6 Jahre; Kinder mit nicht deutscher Muttersprache: 3 %; 444 Grundschulen). In der Sekundarstufe in der 8. Klasse verteilten sich die Jugendlichen auf 93 Gymnasien (n = 4 776) und 248 Schulen (nachfolgend als SMBG bezeichnet; n = 6 163; Details Tab. ESM-4), die mehrere Bildungsgänge anbieten.

[4] Um keine neuartigen Vergleiche zu ermöglichen, wird das Bundesland nicht benannt.

Messinstrumente

In der Hauptstudie umfasste der VERA-3-Test in Mathematik insgesamt 45 Items (mit offenen und geschlossenen Antwortformaten), die zwei inhaltsbezogene Kompetenzen (Leitideen) erfassen: Zahlen und Operationen sowie Muster und Strukturen. Der VERA-8-Test umfasste insgesamt 46 Items (mit offenen und geschlossenen Antwortformaten), die sich auf alle fünf Leitideen bezogen: Zahlen und Operationen, Muster und Strukturen, Raum und Form, Größen und Messen sowie Daten, Häufigkeit und Wahrscheinlichkeit (Beispielaufgaben s. Institut zur Qualitätsentwicklung im Bildungswesen [IQB], n. d.a). Die Reliabilität (interne Konsistenz, KR-20) lag in der 3. und 8. Klasse bei jeweils .90. Der VERA-3-Test zu Deutsch-Lesen umfasste 20 Items (KR-20 = .75), der entsprechende VERA-8-Test 36 Items (KR-20 = .83). Für die Durchführung aller Tests sowie deren Korrektur (auf Basis eines Kodiermanuals) und die Dateneingabe waren Lehrkräfte zuständig. Die SuS bearbeiteten die Tests im Fach Deutsch und Mathematik jeweils an separaten Tagen innerhalb des zweiten Schulhalbjahres mit einem zeitlichen Abstand von einem Tag bzw. fünf Tagen. Zudem gaben Lehrkräfte für die SuS zu beiden Klassenstufen die Halbjahresnote in Mathematik und Deutsch auf der Notenskala von 1 (*sehr gut*) bis 6 (*ungenügend*) an (Interkorrelationstabelle s. ESM-4).

Datenanalyse

Skalierung der bildungsstandardbasierten Tests

Die Skalierung aller Testleistungen erfolgte auf Basis eines eindimensionalen Rasch-Modells (Rasch, 1980) für dichotome Daten. Hierzu wurden die Itemparameter aus den Pilotierungsstudien des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) herangezogen. Die Personenparameter wurden für die nachfolgenden Analysen unter Nutzung der Software R (Paket: sirt) auf die Bildungsstandardmetrik transformiert, um die erreichten Kompetenzstufen zu bestimmen.

Umgang mit fehlenden Werten

In der Hauptstudie lag für die Analysestichprobe der Anteil fehlender Werte pro Variable bei maximal 11 % (Tab. ESM-4). Wir nutzten das multiple Imputationsverfahren MICE (Buuren & Groothuis-Oudshoorn, 2011), um jeweils 15 vollständige Datensätze zu erzeugen. Bei der

Imputation der fehlenden Werte berücksichtigten wir das Skalenniveau sowie die hierarchische Datenstruktur (Schulebene). Um die Qualität der imputierten Daten zu verbessern (Collins, Schafer & Kam, 2001), nutzten wir neben den Prädiktoren und Kriterien folgende Hilfsvariablen: Alter, Geschlecht, nicht deutsche Muttersprache und das Vorliegen besonderer Lernschwierigkeiten im Verhalten, Rechnen und Schriftspracherwerb.

Statistische Analysemodelle: klassifikatorischer Ansatz

Zur Bewertung der Klassifikationsgüte nutzten wir zwei einschlägige Indizes: Sensitivität und positiv prädiktiver Wert (Bossuyt et al., 2015; Petscher, Kim & Foorman, 2011; für weitere Indizes s. Diskussion, ESM-6 bis ESM-26). Diese Indizes bieten komplementierende Informationen, wie gut Tests, Noten oder eine Kombination beider Prädiktoren SuS identifizieren können, die „gefährdet" sind, zukünftige Bildungsergebnisse zu verfehlen.

Bemessen an der Gesamtanzahl aller SuS, die spätere Bildungsergebnisse tatsächlich verfehlten, gibt die Sensitivität (*Sen*) den prozentualen Anteil von SuS an, die richtig als „gefährdet" identifiziert wurden: *Sen* = RP / (RP + FN) mit RP als der Anzahl Richtig Positiver und FN als der Anzahl Falsch Negativer. Bemessen an der Gesamtanzahl aller SuS, die in der Grundschule als „gefährdet" klassifiziert wurden, gibt der positiv prädiktive Wert (*ppW*) den prozentualen Anteil von SuS an, die die späteren Bildungsergebnisse tatsächlich verfehlten: *ppW* = RP / (RP + FP) mit FP als der Anzahl Falsch Positiver.

Die Güte der Prädiktoren für das Verfehlen von späteren Bildungsstandards bzw. Noten untersuchten wir jeweils getrennt für die Fächer Mathematik und Deutsch bzw. getrennt für SuS an Gymnasien und SMBG. Die Indizes und deren 95 %-Konfidenzintervalle (KIs) wurden mit der Software R (Paket: epiR) ermittelt. Insbesondere berücksichtigt epiR die starke Asymmetrie der KIs (u. a. Newcombe & Altman, 2000), sofern Basis- oder Selektionsraten sich den Extremen 0 % bzw. 100 % nähern (dies war vor allem am Gymnasien für N_M/N_D bzw. T_MuN_M/T_DuN_D der Fall). Die Berechnung der KIs erfolgte dabei auf Basis der durchschnittlichen Zellhäufigkeiten (ESM-15 bis ESM-26) der 15 imputierten Datensätze. Eine Bestimmung der KIs auf Basis des üblichen Vorgehens (Rubin, 1987) würde diese Asymmetrie vernachlässigen (s. ESM-27 bis ESM-29). Ein Vergleich der KIs (epiR vs. Formel nach Rubin) zeigte zudem, dass die hier berichteten KIs ein konservativeres Maß der Schätzpräzision für die jeweiligen Indizes darstellen.

Ergebnisse

Die Ergebnisse der Hauptstudie werden in Abb. 1 separat nach Fach (A: Mathematik, B: Deutsch) für Jugendliche an SMBG (s. ESM-30, -31 für die exakten Werte) und für jene an Gymnasien in Abb. 2 (ESM-32, -33) dargestellt. Die Ergebnisse der Ergänzungsstudie mit Grundschulkindern werden in Abb. 3 (ESM-34, -35) visualisiert. In allen Abbildungen markieren die horizontalen Linien bei den *ppW* die jeweilige Basisrate, d. h. den Anteil der SuS, die ein fachbezogenes Bildungskriterium verfehlten. Diese ist eine wichtige Referenzmarke für den *ppW*, da bei niedriger (hoher) Basisrate der *ppW* ebenfalls niedrig (hoch) ist (Petscher et al., 2011).

Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG)

Insgesamt resultierte ein differenziertes Ergebnismuster, das wir für SMBG wie folgt zusammenfassen. (a) Die *Sen* der Testergebnisse in Mathematik und Deutsch (T_M/T_D) lag stets über der der Noten (N_M /N_D) sowohl für eine Prognose des Testkriteriums (Verfehlen des Mindeststandards) als auch des Notenkriteriums (Verfehlen der Noten 1, 2 oder 3). Beispielsweise betrug für T_M/T_D bei einer Orientierung am Verfehlen der Mindeststandards in der 3. Klasse die *Sen* = 42 %/27 % für die Prognose, ob Jugendliche in der 8. Klasse den Mindeststandard in Mathematik/Deutsch (Lesen) verfehlten. Wurde lediglich N_M /N_D in der 3. Klasse zur Identifizierung „gefährdeter“ Kinder herangezogen, wurden deutlich weniger Kinder richtig identifiziert mit *Sen* = 21 %/17 %. Dieses Ergebnismuster zeigte sich für Mathematik ausgeprägter als für Deutsch. (b) Eine Orientierung an den Regelstandards als Schwelle verbesserte stets die *Sen* für T_M/T_D. (c) Eine Kombination von Test und Note mit der ODER-Strategie (T_MoN_M /T_DoN_D) erhöhte stets die *Sen* relativ zu T, N bzw. TuN sowohl im Fach Mathematik als auch Deutsch. Eine Kombination von Test und Note mit der UND-Strategie (T_MuN_M /T_DuN_D) reduzierte jeweils die *Sen* relativ zu T, N bzw. ToN im jeweiligen Fach. (d) Die *ppW* von N_M /N_D waren höher als die für T_M/T_D für beide untersuchten Kriterien in Mathematik und für das Notenkriterium in Deutsch. (e) Eine Orientierung am Regelstandard verminderte die *ppW* für T_M/T_D. (f) Die Kombination TuN erhöhte jeweils die *ppW* gegenüber T und ToN in Mathematik und Deutsch. Höhere Werte für *ppW* gegenüber N zeigten sich mit T_MuN_M /T_DuN_D häufiger mit der Orientierung am Mindeststandard als Schwelle.

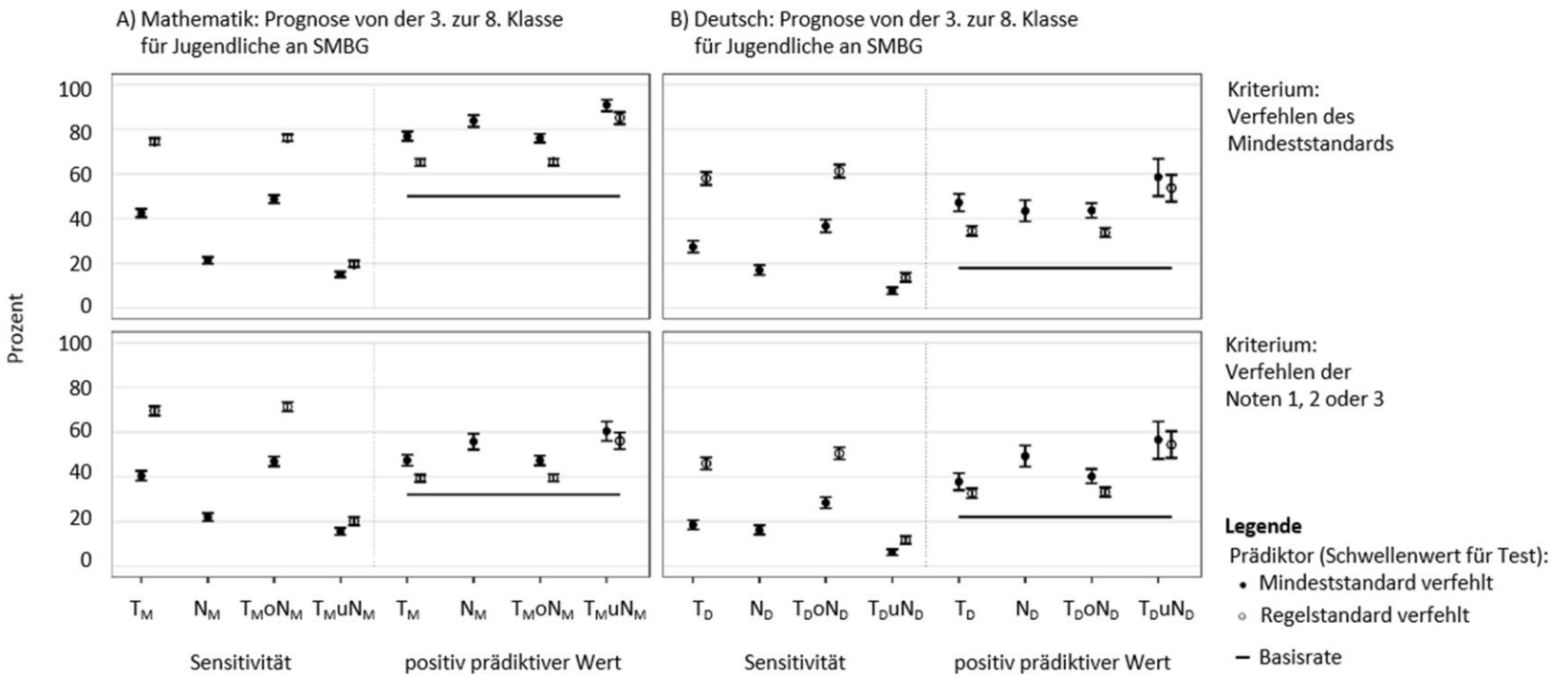


Abbildung 1. Hauptstudie - Ergebnisse für Jugendliche an SMBG in (A) Mathematik und (B) Deutsch: Sensitivität (*Sen*) und positiv prädiktiver Wert (*ppW*, in Prozent) für das Verfehlen des Mindeststandards (= unter Mindeststandard) und das Verfehlen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Strategien (Spezifikation s. Text) und (2) der Schwelle für die Tests (Verfehlen der Mindest- bzw. Regelstandards). T_M bzw. T_D = Testergebnis in Mathematik bzw. Deutsch, N_M bzw. N_D = Note in Mathematik bzw. Deutsch, T_MoN_M bzw. T_DoN_D = Testergebnis oder Note in Mathematik bzw. Deutsch, T_MuN_M bzw. T_DuN_D = Testergebnis und Note in Mathematik bzw. Deutsch. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Jugendlichen, die in der 8. Klasse ein bestimmtes mathematik- bzw. deutschbezogenes Bildungsergebnis verfehlten).

Jugendliche an Gymnasien

Für Jugendliche an Gymnasien sind insbesondere die Ergebnisse der *ppW* für N_M/N_D und T_MuN_M/T_DuN_D mit (großer) Vorsicht zu interpretieren, da aufgrund der geringen Selektionsraten die KIs zwangsläufig relativ weit sind. (a) Es zeigte sich in Mathematik jeweils eine höhere *Sen* für T_M gegenüber N_M für beide Kriterien. Für T_D zeigte sich dies nur bei der Orientierung am Regelstandard als Schwelle. Generell fällt auf, dass die *Sen* für N_M/N_D sehr niedrig ausfällt mit *Sen* $\leq$ 3 %. (b) Eine Orientierung an den Regelstandards als Schwelle verbesserte stets die *Sen* für T_M/T_D. (c) Eine Kombination von Test und Note führte in beiden Fächern weder über die ODER- noch die UND-Strategie zu einer deutlichen Verbesserung oder Verschlechterung der *Sen* gegenüber den Einzelprädiktoren. (d) Generell waren die KIs der *ppW* für N_M/N_D, sowie für T_MuN_M/T_DuN_D sehr weit, sodass wir diese Ergebnisse aufgrund der ungenauen Schätzung der korrespondierenden Populationswerte nicht weiter interpretieren wollen. (e) Eine Orientierung an den Regelstandards als Schwelle verminderte tendenziell den *ppW* für T_M/T_D. Jedoch blieb dieser stets über der Basisrate. (f) Zusammenfassend zeigte sich für Jugendliche an Gymnasien weitestgehend ein vergleichbares Ergebnismuster wie für Jugendliche an SMBG. Jedoch fielen die absoluten Werte für *Sen* und *ppW,* sowie die Unterschiede dieser Kennwerte zwischen den verschiedenen Strategien und Schwellen geringer aus.

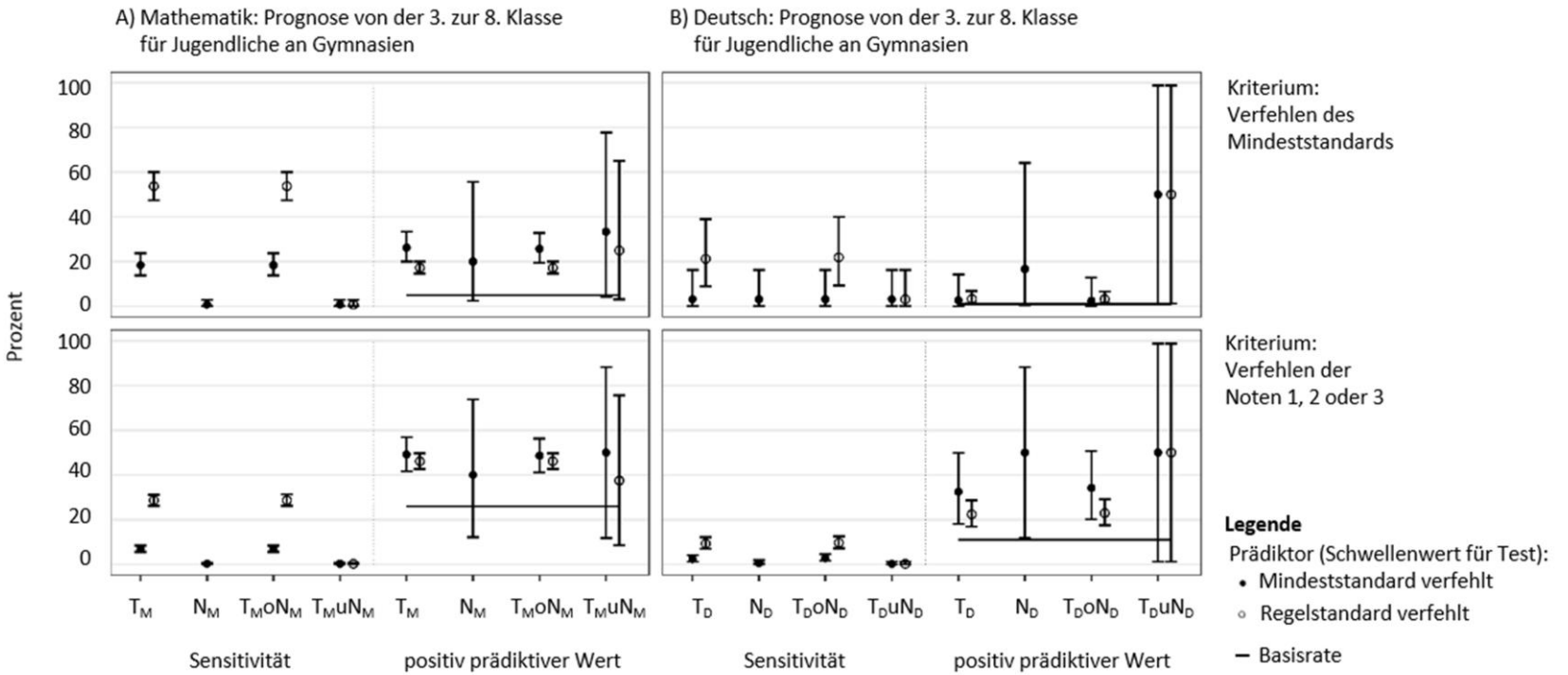


Abbildung 2. Hauptstudie - Ergebnisse für Jugendliche an Gymnasien in (A) Mathematik und (B) Deutsch: Sensitivität (*Sen*) und positiv prädiktiver Wert (*ppW*, in Prozent) für das Verfehlen des Mindeststandards (= unter Mindeststandard) und das Verfehlen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Strategien (Spezifikation s. Text) und (2) der Schwelle für die Tests (Verfehlen der Mindest- bzw. Regelstandards). T_M bzw. T_D = Testergebnis in Mathematik bzw. Deutsch, N_M bzw. N_D = Note in Mathematik bzw. Deutsch, $T_M o N_M$ bzw. $T_D o N_D$ = Testergebnis oder Note in Mathematik bzw. Deutsch, $T_M u N_M$ bzw. $T_D u N_D$ = Testergebnis und Note in Mathematik bzw. Deutsch. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Jugendlichen, die in der 8. Klasse ein bestimmtes mathematik- bzw. deutschbezogenes Bildungsergebnis verfehlten).

Ergänzungsstudie: Prognose von der 4. zur 5. bzw. 6. Klasse in der Grundschule

In der Ergänzungsstudie untersuchten wir, ob sich für das Fach Mathematik die Ergebnisse der Hauptstudie innerhalb der Primarstufe replizieren lassen. Insgesamt (Abb. 3) zeigten sich (a) für die Vorhersage beider Kriterien vergleichbar hohe *Sen* für T_M wie für Jugendliche an SMBG. T_M wies mit der Orientierung am Mindeststandard als Schwelle in etwa vergleichbare *Sen* auf wie für N_M. Eine Orientierung am Regelstandard hatte zur Folge, dass die *Sen* von T_M höher war als die von N_M. Im Vergleich zu den *Sen*, die üblicherweise für Testanwendungen im deutschsprachigen Raum gefunden werden (grau hinterlegte Werte in Abb. 3, Details s. ESM-36), liegen die *Sen* für T_M mit Orientierung am Mindeststandard tendenziell darunter und mit Orientierung am Regelstandard tendenziell darüber. (b) Es zeigten sich ebenso wie für SMBG weitestgehend höhere *ppW* der N_M gegenüber T_M. Im Vergleich zu den *ppW* anderer Schulleistungstests, liegen die *ppW* für T_M mit Orientierung am Mindeststandard vergleichbar hoch bzw. sogar höher und mit Orientierung am Regelstandard tendenziell niedriger. (c) Für die beiden Kombinationsstrategien von Test und Note, ließen sich die Befundmuster für *Sen* und *ppW* für SMBG weitestgehend replizieren. So lässt sich durch $T_M o N_M$ jeweils die *Sen* gegenüber T, N bzw. TuN optimieren. Damit wurden bei einer Orientierung am Mindeststandard vergleichbare und mit Orientierung am Regelstandard sogar meist höhere *Sen* im Vergleich zu anderen Schulleitungstests erreicht. $T_M u N_M$ verbesserte jeweils die *ppW* gegenüber T, N bzw. ToN. Mit dieser Strategie wurden auch *ppW* erreicht, die über denen anderer Schulleistungstests für den deutschsprachigen Raum liegen.

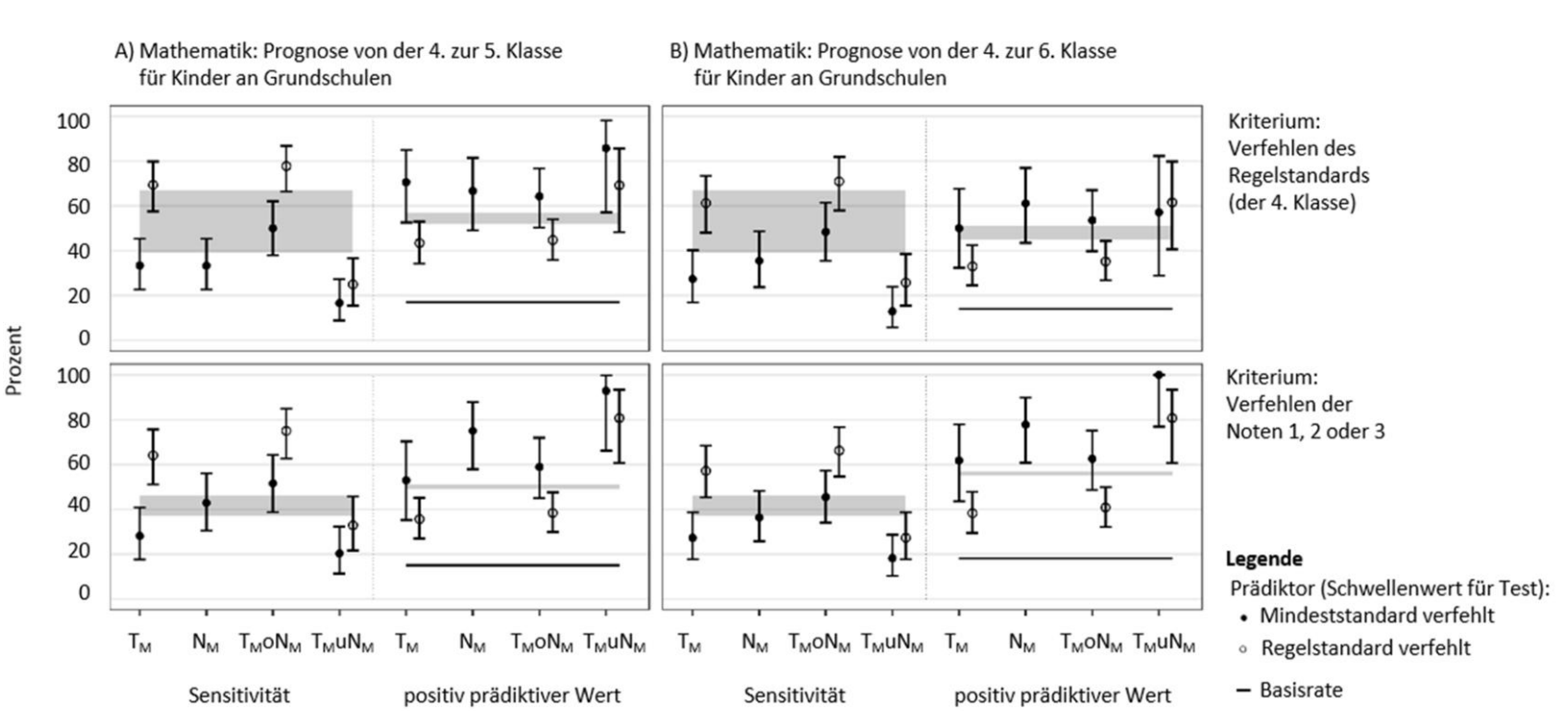


Abbildung 3. Ergänzungsstudie - Ergebnisse für Kinder an Grundschulen im Fach Mathematik für (A) Prognosen von der 4. zur 5. Klasse und (B) Prognosen von der 4. zur 6. Klasse: Sensitivität (*Sen*) und positiv prädiktiver Wert (*ppW*, in Prozent) für das Verfehlen des Regelstandards (= unter Regelstandard) und das Verfehlen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Strategien (Spezifikation s. Text) und (2) der Schwelle für die Tests (Verfehlen der Mindest- bzw. Regelstandards). T_M = Testergebnis in Mathematik, N_M = Note in Mathematik, $T_M oN_M$ = Testergebnis oder Note in Mathematik, $T_M uN_M$ = Testergebnis und Note in Mathematik. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Kinder, die in der 5. bzw. 6. Klasse ein bestimmtes mathematikbezogenes Bildungsergebnis verfehlten). Die grau unterlegten Werte markieren den Referenzbereich, der für deutschsprachige standardisierte Schulleistungstests ermittelt wurde (für Details s. ESM 36).

Diskussion

In diesem Beitrag untersuchten wir im Rahmen von zwei Längsschnittstudien erstmals aus einer klassifikatorischen Perspektive, inwiefern die Ergebnisse von Bista-Tests in Mathematik und Deutsch-Lesen in der 3. und 4. Klasse geeignet sind, Kinder zu identifizieren, die „gefährdet“ sind, bis zu 5 Jahre spätere Bildungsergebnisse (Testleistungen und Schulnoten) zu verfehlen.

Wie gut sind bildungsstandardbasierte Tests zur Identifizierung „gefährdeter“ Kinder geeignet?

Wenn Kinder die Mindeststandards in der 3. Klasse verfehlten, zeigten die *Sen* für die VERA-Tests in Mathematik bzw. Deutsch (Lesen), dass 18 bis 42 % der Jugendlichen an SMBG und 2 bis 18 % am Gymnasium richtig als „gefährdet“ identifiziert wurden, wichtige Bildungsergebnisse (Mindeststandard; Noten 1 bis 3) zu verfehlen. Bemessen an der Gesamtanzahl aller Kinder, die in der 3. Klasse mit den VERA-Tests als „gefährdet“ klassifiziert wurden, hatten an SMBG *ppW* = 38 bis 77 % und an Gymnasien *ppW* = 3 bis 49 % die späteren Bildungsergebnisse auch tatsächlich verfehlt. In der Ergänzungsstudie war die Güte des bildungsstandardbasierten Mathematiktests in der 4. Klasse (*Sen* = 27 – 33 %, *ppW* = 50 – 71 %) vergleichbar mit der für Jugendliche an SMBG. Dieses Ergebnis steht im Einklang mit der Studie von Hoffmann und Böhme (2017) auf der Basis von Querschnittsdaten für Kinder in der 4. Klasse im Fach Deutsch.

Identifizierte man Kinder auf der Grundlage einer höheren Schwelle, und zwar durch das Verfehlen des Regelstandards, zeigte sich zum einen bei fast allen Kriterien mindestens eine Verdopplung der *Sen* im Vergleich zur Orientierung am Mindeststandard (SMBG: *Sen* = 46 – 75 %; Gymnasium: *Sen* = 9 – 54 %; Grundschule: *Sen* = 57 – 69 %). Zum anderen nahm der *ppW* ab, der jedoch auch dann nicht unter die Basisrate sank.

Wie ist die Klassifikationsgüte bildungsstandardbasierter Tests im Vergleich zu jener von Schulnoten zu bewerten?

Im Rahmen unserer Studien zogen wir die Schulnoten als praxisrelevanten Vergleichsmaßstab für die Klassifikationsgüte von Bista-Tests heran. Es zeigte sich tendenziell eine (deutlich) höhere *Sen* für die Tests der 3. Klasse als für die Noten der 3. Klasse. Jedoch zeichnete sich

für die Identifizierung der „gefährdeten“ Kinder auf Basis der Note tendenziell ein höherer *ppW* ab, zumindest für Jugendliche an SMBG. Orientiert man sich bei der Identifizierung am Regelstandard, sind der Zugewinn an der *Sen* wie auch der Verlust am *ppW* der Tests gegenüber den Noten noch ausgeprägter. In der Ergänzungsstudie zeigte sich bei der Identifizierung in der 4. Klasse am Regelstandard weitestgehend dasselbe Muster. Für die Identifizierung am Mindeststandard schneidet der Test in der *Sen* jedoch eher gleich bis leicht schlechter ab im Vergleich zur Note. Für die *ppW* deutete sich erneut der Vorteil der Noten gegenüber den Tests an.

Eine Erklärung für die insgesamt besseren *ppW* der (Halbjahres-)Noten gegenüber den Tests könnte sein, dass es sich bei diesen Zeugnisnoten im Gegensatz zu den Tests nicht um eine einmalige Messung, sondern um ein aggregiertes Leistungsmaß handelt, in das zahlreiche (schriftliche und mündliche) Leistungsbeobachtungen über einen längeren Zeitraum eingingen. Möglicherweise wird dieser Effekt jedoch bei den *Sen* der Noten davon überlagert, dass die Selektionsraten der Zeugnisnoten 4, 5 und 6 niedrig sind, was (ceteris paribus) die *Sen* verringert.

Einordnung der Befunde in den Forschungsstand

Generell ist festzustellen, dass alle ermittelten *Sen* von Bista-Tests unter den Richtwerten (*Sen* $\geq$ 70 %) lagen, die im englischsprachigen Raum für die *Sen* von Screenings gefordert werden (Kilgus, Methe, Maggin & Tomasula, 2014). Mit diesen *Sen* nehmen Bista-Tests jedoch keine Sonderrolle ein, denn mit einem Median der *Sen* von 56 % liegen auch die meisten etablierten schulischen Mathematiktests deutlich unterhalb des Richtwertes (ESM-1). Für Deutschtests liegt der Median der *Sen* mit 65 % knapp unter dem Richtwert.

Weiterhin gilt für Indizes zur Bewertung der Klassifikationsgüte wie bspw. *Sen* und *ppW*, dass diese die „Leistungsfähigkeit eines Verfahrens unter bestimmten Anwendungsbedingungen [charakterisieren], die durch die Basis- und Selektionsraten bestimmt werden“ (Tröster, 2009, S. 139). Dies gilt insbesondere für den *ppW* (Petscher et al., 2011; Streiner, 2003), aber auch für die *Sen* (Kilgus et al., 2014; Whiting, Rutjes, Westwood & Mallett, 2013). Zudem gestaltet sich vor allem die Einordnung der Befunde der Hauptstudie in den bisherigen Forschungsstand schwierig, da es für den deutschsprachigen Raum keine Vorgängerstudie mit vergleichbarem Prognosezeitraum sowie vergleichbaren Basis- und Selektionsraten gibt (s. ESM-36.1). Insbesondere bei längeren Prognosezeiträumen

ist von einer „validity degradation“ auszugehen: Längere Prognosezeiträume gehen in der Regel mit einer geringeren Prognosegenauigkeit einher (Dahlke, Kostal, Sackett & Kuncel, 2018; Sen: Kilgus et al., 2014). Lediglich für die Ergänzungsstudie gibt es Vorgängerstudien, in denen die relevanten Parameter (SR, BR, Prognosezeitraum) weitestgehend vergleichbar waren (s. ESM-36.2). Relativ zu anderen Schulleistungstests in Mathematik resultierten für Bista-Tests geringere *Sen* bei einer Orientierung am Mindeststandard, aber meist höhere oder zumindest vergleichbare *Sen* bei einer Orientierung am Regelstandard. Für *ppW* existieren keine Richtwerte. Diese sind direkt von der Höhe der Basisrate beeinflusst und somit noch weniger zwischen Studien vergleichbar (u. a. Petscher et al., 2011; Streiner, 2003). Um hier dennoch Vergleiche anstellen zu können, schätzten wir für die Ergänzungsstichprobe mit der Methode von Taylor und Russel (1939) die *ppW* auf Grundlage der Basisraten unserer Studie sowie Parametern, die typisch für Schulleistungstests sind (Details s. ESM-36). Für die Ergänzungsstudie zeigte sich, dass mit den Bista-Mathematiktests mit der Orientierung am Mindeststandard meist vergleichbare oder sogar höhere *ppW* erreicht wurden als mit anderen Schulleistungstests. Zusammenfassend deuten die vorliegenden Ergebnisse also darauf hin, dass Bista-Tests – in Abhängigkeit der gewählten Schwelle – vergleichbare (wenn nicht sogar bessere) *Sen* und *ppW* aufweisen können als alternative etablierte Schulleistungstests.

Nutzen für die pädagogisch-psychologische Diagnostik an Schulen

Lehrkräfte aller öffentlichen Schulen sind zur Durchführung von VERA verpflichtet. Sie gelangen so an Informationen zu ihren SuS, die eine Vorhersage über Noten hinaus (deutlich) verbessern können. Unsere Befunde veranschaulichen, dass verschiedene Strategien zur Kombination von Testergebnis und Note bzw. die gewählte Schwelle sich darin unterscheiden, wie gut bestimmte diagnostische Zielsetzungen erreicht werden können. In beiden Studien konnte für die Prognosen die höchste *Sen* beobachtet werden, wenn das Testergebnis mit der Note über eine ODER-Strategie kombiniert wurde. Der *ppW* blieb unabhängig von der gewählten Strategie stets mindestens auf dem Niveau, wie er separat für die Tests ermittelt wurde. Insgesamt zeigen diese Befunde, dass durch die Kombination beider Prädiktoren die diagnostische Entscheidung verbessert werden kann (oder sich zumindest nicht verschlechtert). Folglich liefern Bista-Tests und Noten Informationen für die Diagnostik, die sich wechselseitig ergänzen können. Die gewählte Schwelle und vor allem die gewählte Kombinationsstrategie von Tests und Noten wirkt sich jedoch unterschiedlich auf die Klassifikationsgüte aus: Wenn Lehrkräfte Bista-Tests lediglich als Bestätigung des bereits

vorliegenden diagnostischen Wissens berücksichtigen (UND-Strategie), wie dies bspw. Lehrkräfte im Kontext von VERA berichtet haben (strategische Nutzung, Maier, 2009), wird Potenzial verschenkt, mehr „gefährdete“ Kinder korrekt zu identifizieren bzw. in Kauf genommen, diese zu übersehen. Im Gegenzug werden weniger Kinder fälschlicherweise als „gefährdet“ eingestuft. Somit ist die UND-Strategie dann empfehlenswert, wenn bspw. eine Auswahl von Kindern für eine vom Umfang limitierte Fördermaßnahme zu treffen wäre. Nutzt man jedoch die Informationen der Tests als ergänzenden Hinweis (ODER-Strategie; instrumentelle Nutzung) wird das Potenzial zur korrekten Identifizierung von mehr „gefährdeten“ Kindern genutzt: Diese Kinder könnten durch entsprechende pädagogische Maßnahmen frühzeitig in ihrer Kompetenzentwicklung gefördert werden. Eine Favorisierung dieser Strategie geht jedoch auf Kosten eines höheren Anteils an fehlklassifizierten „gefährdeten“ Kindern. Welche diagnostische Strategie letztlich favorisiert wird, sollte von der Antwort auf folgende Fragen abhängig gemacht werden: Was bedeutet es, wenn Kinder korrekt (fälschlicherweise) als förderwürdig diagnostiziert werden? Stehen genügend zeitliche und personelle Ressourcen für die Förderung zur Verfügung? Hierbei ist anzumerken, dass zur Ableitung diagnostischer Strategien auch die Spezifität und der negativ prädiktive Wert genutzt werden können (s. ESM-6 bis ESM-14). Diese Indizes liefern komplementierende Informationen, wie gut Tests, Noten oder beide Prädiktoren SuS identifizieren können, die „nicht gefährdet“ sind, die von uns definierten Bildungsergebnisse zu verfehlen. Hier zeigte sich ein einheitliches Befundmuster: Abgesehen von einer Ausnahme wurde mit allen Prädiktoren bzw. Strategien, für beide Schulformen und Fächer, der verlangte Richtwert für Screenings von Spezifität ≥ 70 % erreicht (Kilgus et al., 2014). Zudem deuten die Befunde darauf hin, dass die Spezifitäten und negativ prädiktive Werte vergleichbar sind mit denen etablierter Schulleistungstests.

Grenzen der Studie

Die vorliegende Studie weist mehrere Grenzen auf, die die Generalisierbarkeit der Ergebnisse einschränken können. Erstens, zur Erfassung aller fachspezifischen Kompetenzen werden jedes Schuljahr neuzusammengestellte VERA-Tests eingesetzt. Es gibt dabei keine Überlappung der eingesetzten Items zwischen VERA-Tests verschiedener Schuljahre; eine schuljahrübergreifende Metrik wird auf Basis von Item-Response-Modellen im Rahmen von vorgeschalteten Studien etabliert. Dabei ist insbesondere zu bedenken, dass mit den VERA-3-Tests in Mathematik in jedem Schuljahr lediglich zwei (der insgesamt fünf)

inhaltsbezogenen Leitideen geprüft werden. Im nachfolgenden Schuljahr wird eine Leitidee durch eine der übrigen drei Leitideen ersetzt (u. a. IQB, n. d.b). Winkelmann und Robitzsch (2009) konnten aber zeigen, dass die inhaltsbezogenen Kompetenzen einen erheblichen Anteil gemeinsamer Varianz besitzen und schlussfolgerten, dass diese ohne großen Informationsverlust zu einem globalen Kompetenzstufenmodell (wie wir es auch in dieser Studie verwendeten) kombiniert werden können. In zukünftigen Studien ist allerdings empirisch zu prüfen, inwiefern sich die vorliegenden Befunde, die in Mathematik und in Deutsch (Lesen) jeweils auf einer bestimmten Itemauswahl basieren, auf andere Itemkombinationen (z. B. Items zur Messung weiterer Leitideen) generalisieren lassen, zumal die Bundesländer bei VERA in einem engen Rahmen auch die Itemauswahl an die spezifischen Gegebenheiten der jeweiligen Bildungssysteme anpassen können.

Zweitens, können sich die Durchführungs- und Auswertungsbedingungen der Tests systematisch auf die Klassifikationsgüte der VERA-Tests auswirken. Leistungsunterschiede bei VERA-8-Tests können (müssen aber nicht zwangsläufig) aus der Durchführung und Auswertung durch geschulte Testleitungen einerseits und durch Lehrkräfte andererseits resultieren (Graf, Emmrich, Harych & Brunner, 2013; Spoden, Fleischer & Leutner, 2014). Eine geringere Objektivität, die bei einer Durchführung und Auswertung der VERA-Tests durch Lehrkräfte resultieren kann, würde die Reliabilität und damit auch die prognostische Validität der Testergebnisse abschwächen. Da in der Hauptstudie Lehrkräfte die Durchführung und Auswertung übernahmen, sind die vorliegenden Ergebnisse eher als eine Untergrenze für die Klassifikationsgüte der Tests zu werten. Für die Ergänzungsstudie trifft diese Einschränkung nicht zu, da hier geschulte Testleitungen die Tests durchführten und auswerteten.

Drittens ist bislang nicht geklärt, wie die Verwendung der VERA-Testergebnisse (z. B. durch die Schulaufsicht oder -inspektion) die Durchführungs- und Auswertungsqualität durch die Lehrkräfte und damit auch die Klassifikationsgüte der VERA-Tests beeinflusst. Aufgrund der Erfahrungen im angloamerikanischen Raum, in denen viele Kompetenzmessungen im Schulbereich „high-stakes“-Charakter haben (Amrein-Beardsley, Berliner & Rideau, 2010), ist zu erwarten, dass Lehrkräfte den ihnen zur Verfügung stehenden Spielraum zur Beeinflussung der Ergebnisse durch intendierte und nicht intendierte Handlungen umso mehr nutzen, je stärker sie VERA als ein Kontrollinstrument (bspw. der Schuladministration) wahrnehmen und sie für die VERA-Ergebnisse ihrer SuS Rechenschaft ablegen müssen.

Viertens, die Zusammensetzung der Schülerschaft (z. B. das durchschnittliche Leistungsniveau) beeinflusst die Klassifikationsgüte von Bista-Tests (Petscher et al., 2011). So sind zum Beispiel für Schulen einer bestimmten Schulform mit höheren (niedrigeren) Basisraten, die durch ein geringeres (höheres) Leistungsniveau der Schülerschaft entstehen, ein höherer (niedrigerer) *ppW* zu erwarten als wir in dieser Arbeit berichten (Streiner, 2003). Diese Überlegung spiegelt sich auch in den Ergebnissen der vorliegenden Hauptstudie wider: Für Jugendliche an Gymnasien ist die Basisrate deutlich geringer als an SMBG. Dies erschwert die Klassifikation, was (relativ zu SMBG) mit deutlich niedrigeren *Sen* und *ppW* einhergeht. Dies impliziert, dass die vorliegenden Ergebnisse für die Schülerschaft an einer bestimmten Schulform als Ganzes gelten, die Güte für Einzelschulen (aufgrund einer abweichenden Basisrate) jedoch sowohl über als auch unter den berichteten Werten liegen kann. Eine Möglichkeit zur Abschätzung des Nutzens verschiedener Screeningverfahren auf der Basis kontextueller Faktoren an Schulen (wie der Basisrate) liefert zum Beispiel VanDerHeyden (2013).

Schlussfolgerung

Lehrkräften gelingt es nur bedingt, die Leistungen ihrer SuS über einen klasseninternen Referenzrahmen hinaus einzuschätzen (Brunner, Anders, Hachfeld & Krauss, 2011; Nachtigall, 2013). Vor allem in leistungsstärkeren Gruppen gelingt Lehrkräften die Identifizierung von „gefährdeten“ SuS weniger gut (VanDerHeyden & Witt, 2005). Die Ergebnisse von VERA-Tests können genutzt werden, um als Screeningverfahren die pädagogisch-psychologische Diagnostik an Schulen grundlegend zu verbessern, da sie flächendeckend in Deutschland in der 3. und 8. Klasse eingesetzt werden: VERA-Tests haben (vor allem in Kombination mit Schulnoten) das psychometrische Potenzial, SuS zu identifizieren, die „gefährdet“ sind, wichtige Bildungsergebnisse zu verfehlen. Wir raten aber stark davon ab, VERA-Testergebnisse für eine weitreichende Selektionsdiagnostik (z. B. im Rahmen der Schullaufbahnempfehlung) zu verwenden (IQB, n. d.c). Jedoch empfehlen wir, aus den VERA-Testergebnissen in Kombination mit Schulnoten (und ggf. weiteren lernrelevanten Informationen), diagnostische Entscheidungen zum Förderbedarf von SuS abzuleiten. Die getroffenen Entscheidungen sollten dann auf Grundlage einer lernbegleitenden Diagnostik (z. B. durch weitere Tests und Beobachtungen im Unterricht) fortlaufend überprüft werden.

Literatur

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Amrein-Beardsley, A., Berliner, D. C. & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, 18* (14), 1–36.

Bos, W., Tarelli, I., Bremerich-Vos, A. & Schwippert, K. (Hrsg.). (2012). *IGLU 2011: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann. Verfügbar unter: http://www.waxmann.com/fileadmin/media/zusatztexte/2828Volltext.pdf

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. et al. (2015). STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Radiology*, *277* (3), 826–832.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Münster: Waxmann.

Buuren, S. van & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal Of Statistical Software, 45* (3), 1–67.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6* (4), 330–351.

Dahlke, J. A., Kostal, J. W., Sackett, P. R. & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. Advance online publication. *Journal of Applied Psychology*.

Fuchs, G. & Brunner, M. (2017). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie, 31* (1), 27–39.

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K. & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78* (4), 423–445.

Graf, T., Emmrich, R., Harych, P. & Brunner, M. (2013). Durchführungseffekte bei Vergleichsarbeiten in Jahrgangsstufe 8. *Empirische Pädagogik*, *27* (4), 459–473.

Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie*, *30* (4), 201–211.

Hellrung, K. & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F.E. Weinert (Hrsg.), *Enzyklopädie der Psychologie: Psychologie des Unterrichts und der Schule* (Band 3, S. 71–176). Göttingen: Hogrefe.

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe*. Münster: Waxmann.

Hoffmann, L. & Böhme, K. (2017). Wird sprachlicher Förderbedarf in der Grundschule sicher erkannt? Zur Klassifikationsgüte von diagnostischen Entscheidungen. *Zeitschrift für Pädagogische Psychologie*, *31* (2), 137–147.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik*. Weinheim Basel: Beltz Verlag.

Institut zur Qualitätsentwicklung im Bildungswesen. (n. d.a). VERA-Aufgabenpool. Verfügbar unter: https://www.iqb.hu-berlin.de/vera/aufgaben

Institut zur Qualitätsentwicklung im Bildungswesen. (n. d.b). VERA-3 Testdomänen 2006 bis 2020 und VERA-8 Testdomänen 2009 bis 2020. Verfügbar unter: https://www.iqb.hu-berlin.de/vera/aktuell

Institut zur Qualitätsentwicklung im Bildungswesen. (n. d.c). 13: Ist eine Schullaufbahnentscheidung durch VERA möglich?. Verfügbar unter: https://www.iqb.hu-berlin.de/vera/faq/#faq13

Kilgus, S. P., Methe, S. A., Maggin, D. M. & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52* (4), 377–405.

Klauer, K. J. (1987). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.

Köller, O. & Reiss, K. (2013). Mathematische Kompetenzen messen: Gibt es Unterschiede zwischen standardisierten Verfahren und diagnostischen Tests? (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, A. Heinze, W. Schneider & U. Trautwein (Hrsg.), *Diagnostik mathematischer Kompetenzen* (Band 11, S. 25–37). Göttingen: Hogrefe.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, *22* (2), 18–26.

Kultusministerkonferenz. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Verfügbar unter: https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/Beschluesse_Veroeffentlichungen/Bildungsmonitoring_Broschuere_Endf.pdf

Kultusministerkonferenz. (2012). *Vereinbarung zur Weiterentwicklung von VERA*. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Kultusministerkonferenz. (2013a). *Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Primarbereich (Jahrgangsstufe 4)*. Verfügbar unter: http://www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Mathemati_2.pdf

Kultusministerkonferenz. (2013b). *Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich „Lesen – mit Texten und Medien umgehen" – Primarbereich.* Verfügbar unter: https://www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Deutsch_L_2.pdf

Kultusministerkonferenz. (2013c). *Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Mathematik.* Verfügbar unter: https://www.iqb.hu-berlin.de/bista/ksm

Kultusministerkonferenz. (2014). *Integriertes Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Deutsch für den Kompetenzbereich Lesen – mit Texten und Medien umgehen.* Verfügbar unter: https://www.iqb.hu-berlin.de/bista/ksm

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: Verlag für Sozialwissenschaften.

Maier, U. (2009). Testen und dann? Ergebnisse einer qualitativen Lehrerbefragung zur diagnostischen Funktion von Vergleichsarbeiten. *Empirische Pädagogik*, *23*, 191–207.

Marx, H. (1992). Frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten: Bestandsaufnahme bisheriger Bewertungsgesichtspunkte von Längsschnittstudien. *Zeitschrift für Pädagogische Psychologie*, *6* (1), 35–48.

Nachtigall, C. (2013). *Landesbericht. Thüringer Kompetenztests 2013*. Verfügbar unter: https://www.kompetenztest.de/downloads/kompetenztests

Newcombe, R. G. & Altman, D. G. (2000). Proportions and their differences. In D.G. Altman, D. Machin, T. Bryant & M. Gardner (Hrsg.), *Statistics with Confidence* (S. 45–56). Bristol: BMJ Books.

Petscher, Y., Kim, Y.-S. & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention*, *36* (3), 158–166.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Edition.). Chicago: University of Chicago Press.

Richter, D. & Böhme, K. (2014). Vergleichsarbeiten im Fokus: Welche Funktionen erfüllt der Test aus Sicht von Lehrkräften? *Schulmanagement*, *45* (2), 12–14.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik*, *35* (1), 79–99.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment*, *81* (3), 209–219.

Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, *23* (5), 565–578.

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen.* Göttingen: Hogrefe.

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review*, *42* (4), 402–414.

VanDerHeyden, A. M. & Witt, J. C. (2005). Quantifying context in assessment: Capturing the effect of base rates on teacher referral and a problem-solving model of identification. *School Psychology Review*, *34* (2), 161–183.

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E. & Mallett, S. (2013). A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology*, *66* (10), 1093–1104.

Winkelmann, H. & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule* (S. 169–196). Weinheim: Beltz Verlag.

Elektronische Supplemente (ESM) – Studie III

ESM-1.

Forschungsstand zur prognostischen Klassifikationsgüte bildungsstandardbasierter Tests und standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten.

ESM-2.

Hauptstudie: Deskriptive Statistiken

ESM-3.

Hauptstudie: Analysen zum Stichprobenausfall.

ESM-4.

Hauptstudie: Korrelationsmatrix.

ESM-5.

Details zur Ergänzungsstudie.

ESM-6.

Abbildungen zur Spezifität und zum negativ prädiktiven Wert für SMBG.

ESM-7.

Tabelle zur Spezifität und zum negativ prädiktiven Wert für SMBG mit Orientierung am Mindeststandard in der 3. Klasse.

ESM-8.

Tabelle zur Spezifität und zum negativ prädiktiven Wert für SMBG mit Orientierung am Regelstandard in der 3. Klasse.

ESM-9.

Abbildungen zur Spezifität und zum negativ prädiktiven Wert für Gymnasien.

ESM-10.

Tabelle zur Spezifität und zum negativ prädiktiven Wert für Gymnasien mit Orientierung am Mindeststandard in der 3. Klasse.

ESM-11.

Tabelle zur Spezifität und zum negativ prädiktiven Wert für Gymnasien mit Orientierung am Regelstandard in der 3. Klasse.

ESM-12.

Abbildungen zur Spezifität und zum negativ prädiktiven Wert für Grundschulen.

ESM-13.

Tabelle zur Spezifität und zum negativ prädiktiven Wert für Grundschulen mit Orientierung am Mindeststandard in der 4. Klasse.

ESM-14.

Tabelle zur Spezifität und zum negativ prädiktiven Wert für Grundschulen mit Orientierung am Regelstandard in der 4. Klasse.

ESM-15.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für SMBG mit Orientierung am Mindeststandard in der 3. Klasse im Fach Mathematik.

ESM-16.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für SMBG mit Orientierung am Mindeststandard in der 3. Klasse im Fach Deutsch.

ESM-17.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für SMBG mit Orientierung am Regelstandard in der 3. Klasse im Fach Mathematik.

ESM-18.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für SMBG mit Orientierung am Regelstandard in der 3. Klasse im Fach Deutsch.

ESM-19.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Gymnasien mit Orientierung am Mindeststandard in der 3. Klasse im Fach Mathematik.

ESM-20.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Gymnasien mit Orientierung am Mindeststandard in der 3. Klasse im Fach Deutsch.

ESM-21.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Gymnasien mit Orientierung am Regelstandard in der 3. Klasse im Fach Mathematik.

ESM-22.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Gymnasien mit Orientierung am Regelstandard in der 3. Klasse im Fach Deutsch.

ESM-23.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Grundschulen mit Orientierung am Mindeststandard in der 4. Klasse. Prognose von der 4. Klasse zur 5. Klasse.

ESM-24.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Grundschulen mit Orientierung am Mindeststandard in der 4. Klasse. Prognose von der 4. Klasse zur 6. Klasse.

ESM-25.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Grundschulen mit Orientierung am Regelstandard in der 4. Klasse. Prognose von der 4. Klasse zur 5. Klasse.

ESM-26.

Tabelle mit weiteren Indizes zur Klassifikationsgüte wie bspw. RATZ-Index für Grundschulen mit Orientierung am Regelstandard in der 4. Klasse. Prognose von der 4. Klasse zur 6. Klasse.

ESM-27.

Abbildungen zur *Sen* und zum *ppW* (Konfidenzintervalle nach Rubin, 1987) für SMBG.

ESM-28.

Abbildungen zur *Sen* und zum *ppW* (Konfidenzintervalle nach Rubin, 1987) für Gymnasien.

ESM-29.

Abbildungen zur *Sen* und zum *ppW* (Konfidenzintervalle nach Rubin, 1987) für Grundschulen.

ESM-30.

Tabelle zur *Sen* und zum *ppW* für SMBG mit Orientierung am Mindeststandard in der 3. Klasse.

ESM-31.

Tabelle zur *Sen* und zum *ppW* für SMBG mit Orientierung am Regelstandard in der 3. Klasse.

ESM-32.

Tabelle zur *Sen* und zum *ppW* für Gymnasien mit Orientierung am Mindeststandard in der 3. Klasse.

ESM-33.

Tabelle zur *Sen* und zum *ppW* für Gymnasien mit Orientierung am Regelstandard in der 3. Klasse.

ESM-34.

Tabelle zur *Sen* und zum *ppW* für Grundschulen mit Orientierung am Mindeststandard in der 4. Klasse.

ESM-35.

Tabelle zur *Sen* und zum *ppW* für Grundschulen mit Orientierung am Regelstandard in der 4. Klasse.

ESM-36.

Details zur Bestimmung von Referenzwerten für die Klassifikationsgüte im Rahmen der Ergänzungsstudie.

Tabelle ESM-1

Forschungsstand zur prognostischen Klassifikationsgüte bildungsstandardbasierter Tests und standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	SR / BR	Sen (in %)	Spe (in %)	ppW (in %)	npW (in %)	RATZ (in %)	LR +	LR -	OR	Quelle
Vorhersage von Mathematiktestergebnis auf Mathematiktestergebnis – kommerziell erhältliche Tests															
DEMAT 1 + (PR < 15)	3	1. → 4.	DEMAT 4 (PR < 15)	< 15	< 15	≈ 1[b]	38	92	24	96	31	4.75[b]	0.67[b]	-	1
	2	1. → 3.	VERA 3 (Vorversion) (PR < 15)	15[b]	10[b]	0.67[b]	28	94	44	88	33[b]	4.67[b]	0.77[b]	5.67[b]	
	2	1. → 3.	DEMAT 3 + (PR < 15)	< 15	< 15	≈ 1[b]	39	93	30	95	32	5.57[b]	0.66[b]	-	
	1	1. → 2.	DEMAT 2 + (PR < 15)	< 15	< 15	≈ 1[b]	44	95	59	92	53	8.80[b]	0.59[b]	-	
DEMAT 2 + (PR < 15)	2	2. → 4.	DEMAT 4 (PR < 15)	< 15	< 15	≈ 1[b]	72	91	34	98	68	8.00[b]	0.31[b]	-	1
	1	2. → 3.	VERA 3 (Vorversion) (PR < 15)	15[b]	13[b]	0.87[b]	45	93	55	90	47	6.72[b]	0.59[b]	11.46[b]	
	1	2. → 3.	DEMAT 3 + (PR < 15)	< 15	< 15	≈ 1[b]	56	91	33	96	49	6.22[b]	0.48[b]	-	
DEMAT 3 + (PR < 15)	1	3. → 4.	DEMAT 4 (PR < 15)	< 15	< 15	≈ 1[b]	52	95	42	97	48	10.40[b]	0.51[b]	-	1
DEZ (n.e.)	≈ 1	0. → 1.	DEMAT 1 + (PR < 15)	6[b]	10[b]	1.67[b]	100	96	60[b]	100[b]	100[b]	24.00[b]	0.00[b]	-	2

(Tabelle ESM-1 wird fortgesetzt)

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	SR/BR	Sen (in %)	Spe (in %)	ppW (in %)	npW (in %)	RATZ (in %)	LR +	LR -	OR	Quelle
ERT 0 + (PR < 16)	≈ 2	1. → 2.	DEMAT 2 + (PR < 16)	18	12	0.67[b]	31 [25;36]	92 [89;95]	47	86[b]	35	3.93[b]	0.75[b]	5.23[b]	3
	≈ 1	1. → 1.	DEMAT 1 + (PR < 16)	16	15	0.94[b]	39 [34;45]	89 [86;93]	42	88[b]	30[b]	3.71[b]	0.68[b]	5.46[b]	
HaReT 1 (= HRT 1) (PR < 16)	≈ 2	1. → 2.	DEMAT 2 + (PR < 16)	19	13	0.68[b]	35 [29;40]	92 [86;97]	49	86[b]	37	4.17[b]	0.71[b]	5.85[b]	3
	≈ 1	1. → 1.	DEMAT 1 + (PR < 16)	16	16	1[b]	49 [44;55]	91 [88;94]	51	90[b]	42	5.39[b]	0.56[b]	9.62[b]	
Kalkulie 1 Diagnoseteil 1 (PR < 16)	≈ 2	1. → 2.	DEMAT 2 + (PR < 16)	19	11	0.58[b]	29 [24;34]	93 [89;96]	47	85[b]	35	3.87[b]	0.77[b]	5.03[b]	3
	≈ 1	1. → 1.	DEMAT 1 + (PR < 16)	16	16	1[b]	52 [46;58]	91 [88;94]	52	91[b]	43	5.70[b]	0.53[b]	10.79[b]	
KR 3-4 (PR = 12)	1	3. → 4.	DEMAT 3 + und 4 (PR < 25)	< 25	< 12	0.48[b]	75	91	-	-	-	8.33[b]	0.27[b]	-	4
(PR = 15)				< 25	< 15	0.6[b]	88	89	-	-	-	8.00[b]	0.13[b]	-	
(PR = 20)				< 25	< 20	0.8[b]	88	86	-	-	-	6.29[b]	0.14[b]	-	
(PR = 24)				< 25	< 24	0.96[b]	88	84	-	-	-	5.50[b]	0.14[b]	-	
(PR = 29)				< 25	< 29	1.16[b]	88	80	-	-	-	4.40[b]	0.15[b]	-	
(PR = 33)				< 25	< 33	1.32[b]	88	77	-	-	-	3.83[b]	0.16[b]	-	
(PR = 39)				< 25	< 39	1.56[b]	100	73	-	-	-	3.70[b]	0.00[b]	-	
(PR = 44)				< 25	< 44	1.76[b]	100	65	-	-	-	2.86[b]	0.00[b]	-	
UGT (n.e.)	≈ 2	1. → 2.	AST 2 / Mathematik (t < 44)	13	10	0.77[b]	36	94	46[b]	91[b]	37	5.48[b]	0.69[b]	7.96[b]	5

(Tabelle ESM-1 wird fortgesetzt)

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	SR/BR	Sen (in %)	Spe (in %)	ppW (in %)	npW (in %)	RATZ (in %)	LR +	LR -	OR	Quelle
Vorhersage von Mathematiktestergebnis auf Mathematiktestergebnis – nicht kommerziell erhältliche Tests															
Testbatterie Dornheim (n.e.)	≈ 2.5	0. → 2.	DEMAT 2 + (PR < 16)	16[b]	25[b]	1.56[b]	60	82	40[b]	91[b]	47	3.40[b]	0.49[b]	7.00[b]	6
	+P (3)			16[b]	28[b]	1.75[b]	70	80	41[b]	93[b]	60[b]	3.57[b]	0.37[b]	9.57[b]	
	≈ 2			16[b]	20[b]	1.25[b]	60	88	50[b]	92[b]	50	5.10[b]	0.45[b]	11.25[b]	
	+P (3)			16[b]	32[b]	2.00[b]	80	77	41[b]	95[b]	72[b]	3.55[b]	0.26[b]	13.74[b]	
	≈ 1.5	0. → 1.	DEMAT 1 + (PR < 18)	18[b]	23[b]	1.28[b]	67	86	52[b]	92[b]	56	4.80[b]	0.38[b]	12.40[b]	
	+P (3)			18[b]	29[b]	1.61[b]	79	82	50[b]	95[b]	72[b]	4.50[b]	0.25[b]	17.80[b]	
	≈ 1			18[b]	23[b]	1.28[b]	67	86	52[b]	92[b]	56	5.19[b]	0.38[b]	13.57[b]	
	+P (3)			18[b]	30[b]	1.67[b]	79	81	48[b]	95[b]	70	4.11[b]	0.26[b]	15.92[b]	
Testbatterie Kaufmann (n.e.)	≈ 2	1. → 2.	AST 2 / Mathematik (t < 44)	15	11	0.73[b]	38	93	50[b]	89[b]	41	5.68[b]	0.67[b]	8.50[b]	5
			Gruppen- und Einzeltest U3 (k.A.)	14	12	0.86[b]	36	92	42[b]	90[b]	32[b]	4.54[b]	0.70[b]	6.51[b]	
M. L. U2 (Einzel- und Gruppentest) (n.e.)	≈ 1	1. → 1.	AST 2 / Mathematik (t < 44)	14	22	1.57[b]	67	86	44[b]	94[b]	57	4.67[b]	0.39[b]	12.00[b]	5
RT U2 (n.e.)	≈ 1	1. → 1.	AST 2 / Mathematik (t < 44)	14	12	0.86[b]	36	92	42[b]	90[b]	32	4.38[b]	0.70[b]	6.27[b]	5
Testbatterie Krajewski (n.e.)	≈ 2.5	0. → 2.	DEMAT 2 + (Vorversion) (PR < 15)	18[b]	11[b]	0.61[b]	47	97	75[b]	89[b]	70	13.74[b]	0.55[b]	25.20[b]	7
	+P(G)			18[b]	16[b]	0.89[b]	53	92	59[b]	90[b]	50	6.54[b]	0.52[b]	12.70[b]	
	≈ 1			14[b]	16[b]	1.14[b]	47	90	43[b]	91[b]	38	4.54[b]	0.59[b]	7.73[b]	
	+P(G)			14[b]	18[b]	1.29[b]	53	88	42[b]	92[b]	43[b]	4.32[b]	0.52[b]	8.02[b]	
	≈ 1.5	0. → 1.	DEMAT 1 + (PR < 15)	15[b]	14[b]	0.93[b]	61	94	65[b]	93[b]	58	10.29[b]	0.41[b]	24.88[b]	
	+P(G)			15[b]	18[b]	1.20[b]	61	89	50[b]	93[b]	52	5.61[b]	0.44[b]	12.86[b]	
	≈ 1			14[b]	18[b]	1.29[b]	65	90	50[b]	94[b]	57[b]	6.35[b]	0.39[b]	16.29[b]	
	+P(G)			14[b]	20[b]	1.43[b]	65	87	45[b]	94[b]	56	5.16[b]	0.40[b]	12.88[b]	

(Tabelle ESM-1 wird fortgesetzt)

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	SR/BR	Sen (in %)	Spe (in %)	ppW (in %)	npW (in %)	RATZ (in %)	LR +	LR -	OR	Quelle
Vorhersage von Mathematiktestergebnis auf Mathematiknote – kommerziell erhältliche Tests															
UGT (n.e.)	1	2. → 2.	Note (Note > 3)	19[b]	9	0.47[b]	18	93	40[b]	83[b]	25[b]	2.76[b]	0.88[b]	3.15 [b]	5
Vorhersage von Mathematiktestergebnis auf Mathematiknote – nicht kommerziell erhältliche Tests															
M. L. U2 (n.e.)	1	2. → 2.	Note (Note > 3)	20	21	1.05[b]	46	85	43[b]	86[b]	31	3.08[b]	0.64[b]	4.80[b]	5
RT U2 (n.e.)	1	2. → 2.	Note (Note > 3)	21	16	0.76[b]	43	91	56[b]	86[b]	45	4.96[b]	0.63[b]	7.93[b]	5
Testbatterie Kaufmann (n.e.)	≈ 2	1. → 2.	Note (Note > 3)	21	11	0.52[b]	48	99	92[b]	88[b]	89	41.13[b]	0.53[b]	77.92[b]	5
VERA 8 (KS = I)	2	8. → 10.	Prüfungsnote Mittlerer Schulabschluss (Note > 4)	6[b]	8[b]	1.33[b]	43[b]	94[b]	30[b]	96[b]	38[b]	7.14[b]	0.60[b]	11.8[b]	8
			(Note > 2)	42[b]	8[b]	0.19[b]	16[b]	98[b]	83[b]	61[b]	70[b]	6.37[b]	0.86[b]	7.39[b]	

(Tabelle ESM-1 wird fortgesetzt)

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	SR/BR	Sen (in %)	Spe (in %)	ppW (in %)	npW (in %)	RATZ (in %)	LR +	LR -	OR	Quelle
Vorhersage von Deutschtestergebnis (im weitesten Sinne Lesen) auf Deutschtestergebnis – kommerziell erhältliche Tests															
BISC (PR < 15)	≈ 2	0. → 2.	Lesefähigkeit (mit Knuspel-L) (k.A.)	14	22	1.57[b]	71	86	44[b]	95[b]	62[b]	4.96[b]	0.33[b]	14.87[b]	9
	≈ 2			15	22	1.47[b]	59	85	42[b]	92[b]	49[b]	4.01[b]	0.48[b]	8.35[b]	
	≈ 1.5			14	14	1.00[b]	62	94	62[b]	94[b]	56	10.21[b]	0.41[b]	25.19[b]	
	≈ 1.5			15	13	0.87[b]	41	93	50[b]	90[b]	39[b]	5.55[b]	0.64[b]	8.69[b]	
	+P (T)			14	17	1.21[b]	76	92	62[b]	96[b]	70[b]	10.06[b]	0.26[b]	39.04[b]	
	+P (T)			15	16	1.07[b]	55	91	52[b]	92[b]	47[b]	6.05[b]	0.50[b]	12.11[b]	
	≈ 1	0. → 1.	Lesefähigkeit (mit Knuspel-L) (k.A.)	15	23	1.53[b]	73	86	47[b]	95[b]	64[b]	5.13[b]	0.32[b]	16.15[b]	
	≈ 1			16	22	1.38[b]	54	84	38[b]	91[b]	40[b]	3.30[b]	0.55[b]	6.02[b]	
	≈ 0.5			15	13	0.87[b]	36	91	40[b]	89[b]	31[b]	3.85[b]	0.70[b]	5.48[b]	
	≈ 0.5			14	13	0.93[b]	32	91	37[b]	89[b]	25[b]	3.45[b]	0.75[b]	4.59[b]	
	+P (T)			15	17	1.13[b]	55	89	46[b]	92[b]	44[b]	4.95[b]	0.51[b]	9.69[b]	
	+P (T)			16	16	1.00[b]	42	88	40[n]	89[b]	30	3.56[b]	0.66[b]	5.38[b]	
DESK 3-6 (k.A.)	1	0. → 1.	DEMAT 1 + (PR ≤ 20)	21	30	1.42[b]	71	81	50	91	58	3.75	0.36	10.43[b]	10
			WLLP (PR ≤ 20)	15	30	2.00[b]	59	75	29	91	41	2.38	0.55	4.35[b]	
HASE (n.e.)	1	3. → 3.	DRT 3	-	-	-	67	60	22	92	42	1.68[b]	0.55[b]	-	11
	1	3. → 3.	WLLP	-	-	-	66	61	27	89	39	1.69[b]	0.56[b]	-	
	1	3. → 3.	Knuspel-L Score 1	-	-	-	82	66	38	93	68	2.41[b]	0.27[b]	-	
	1	3. → 3.	Knuspel-L Score 2 (n.e.)	16	43	2.69	84	64	31	96	72	2.37[b]	0.24[b]	9.72[b]	
PB-LRS (PR < 15)	≈ 1.5	0. → 1.	Rechtschreib-leistung (DBL1) (TW < 9)	10[b]	18	1.80[b]	63	87	36	95[b]	55[b]	4.90[b]	0.43[b]	11.55[b]	12

(Tabelle ESM-1 wird fortgesetzt)

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	SR/BR	Sen (in %)	Spe (in %)	ppW (in %)	npW (in %)	RATZ (in %)	LR +	LR -	OR	Quelle
Vorhersage von Deutschtestergebnis (im weitesten Sinne Lesen) auf Deutschnote – kommerziell erhältliche Tests															
HASE (n.e.)	3	1. → 3.	Deutschnote (Note > 3)	13[b]	44[b]	3.35[b]	83	62	25	96	70	2.20[b]	0.27[b]	8.05[b]	11
			Lesenote (Note > 3)	-	-	-	82	61	19	97	68	2.10[b]	0.30[b]	-	
			Rechtschreibnote (Note > 3)	-	-	-	69	64	36	88	46	1.92[b]	0.48[b]	-	
Vorhersage von Deutschtestergebnis (Lesen) auf Deutschnote – nicht kommerziell erhältliche Tests															
VERA 8 (KS = I)	2	8. → 10.	Prüfungsnote Mittlerer Schulabschluss (Note > 4)	1[b]	2[b]	2[b]	25[b]	98[b]	14[b]	99[b]	23[b]	14.5[b]	0.76[b]	19.06[b]	8
			(Note > 2)	44[b]	2[b]	0.05[b]	5[b]	100[b]	100[b]	58[b]	100[b]	-	0.95[b]	-	

Anmerkungen. SW = Schwellenwert, Dauer = Länge des Vorhersagezeitraums in Jahren, Klassen = Klassen des ersten und zweiten Messzeitpunktes (Vorhersagezeitraum), BR = Basisrate, SR = Selektionsrate, Sen = Sensitivität, Spe = Spezifität, ppW = positiv prädiktiver Wert, npW = negativ prädiktiver Wert, RATZ = RATZ-Index, LR + = positiver Likelihood Ratio, LR - = negativer Likelihood Ratio, OR = Odds Ratio, DEMAT = Deutscher Mathematiktest – für erste, zweite, dritte und vierte Klassen, VERA = Vergleichsarbeiten – für dritte, sechste und achte Klassen (bildungsstandardbasierte Tests), DEZ = Diagnostikum zur Entwicklung des Zahlkonzepts, ERT 0 + = Eggenberger Rechentest, [ ] = 95 %-Konfidenzintervall, HaReT 1 = Hamburger Rechentest für Klasse 1, Kalkulie 1 = Kalkulie Diagnoseaufgaben Teil 1, KR 3-4 = Kettenrechner für 3. und 4. Klassen, UGT = Utrechter Zahlbegriffstest, AST 2 / Mathematik = Allgemeiner Schulleistungstest für 2. Klassen, +P = Zusätzlich wurde zum Testergebnis ein weiteres Testergebnis zur Prognose herangezogen: (3) = Zahlen Lesen, sprachliche Arbeitsgedächtnisleistung, räumliche IQ-Komponente (2 Risikopunkte), M. L. U2 = Mathematischer Leistungstest U2, RT U2 = Rechentest U2, (G) = Gedächtniskapazität, BISC = Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten, Knuspel-L = Knuspels Leseaufgaben, (T) = Testergebnisse zum ersten und zweiten Messzeitpunkt kombiniert, DESK 3-6 = Dortmunder Entwicklungsscreening für den Kindergarten, WLLP = Würzburger Leise-Leseprobe, HASE = Heidelberger Auditives Screening in der Einschulungsuntersuchung, DRT 3 = Diagnostischer Rechtschreibtest für dritte Klassen, PB-LRS = Gruppentest zur Früherkennung von Lese- und Rechtschreibschwierigkeiten, Phonologische Bewusstheit bei Kindergartenkindern und Schulanfängern, DBL1 = Diagnostische Bilderliste.

[a] Schwellenwert (SW) Abkürzungen: PR = Prozentrang, n.e. = Schwellenwert nicht eindeutig angegeben, t = t-Wert, k.A. = keine Angabe, Note = Schulnote, TW = Testwert, KS = Kompetenzstufe. [b] selbst berechnete Werte (Tröster, 2009, S. 105)

Literatur der Tabelle ESM-1

1 Hasselhorn, M., Roick, T. & Gölitz, D. (2005). Stabilitäten und prognostische Validitäten der Mathematikleistungen. Eine Längsschnittstudie mit der DEMAT-Reihe in der Grundschule (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, H. Marx & W. Schneider (Hrsg.), *Diagnostik von Mathematikleistungen* (Band 4, S. 187–198). Göttingen: Hogrefe.

2 Weißhaupt, S., Peucker, S. & Wirtz, M. (2006). Diagnose mathematischen Vorwissens im Vorschulalter und Vorhersage von Rechenleistungen und Rechenschwierigkeiten in der Grundschule. *Psychologie in Erziehung und Unterricht: Zeitschrift für Forschung und Praxis*, (4), 236–245.

3 Gomm, B. (2014). *Prognostische Validität mathematischer Screenings*. Dortmund: Technische Universität Dortmund. Verfügbar unter: https://eldorado.tu-dortmund.de/bitstream/2003/33789/1/Dissertation.pdf

4 Gölitz, D., Roick, T. & Hasselhorn, M. (2013). Kettenrechner für dritte und vierte Klassen (KR 3-4) (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, A. Heinze, W. Schneider & U. Trautwein (Hrsg.), *Diagnostik mathematischer Kompetenzen* (Band 11, S. 149–164). Göttingen: Hogrefe.

5 Kaufmann, S. (2002). *Früherkennung von Rechenstörungen in der Eingangsklasse der Grundschule und darauf abgestimmte remediale Massnahmen* (Europäische Hochschulschriften. Reihe XI, Pädagogik) (Band 880). Frankfurt am Main: Peter Lang.

6 Dornheim, D. (2007). *Prädiktion von Rechenleistung und Rechenschwäche. Der Beitrag von Zahlen-Vorwissen und allgemein-kognitiven Fähigkeiten*. Berlin: Logos Verlag Berlin GmbH.

7 Krajewski, K. (2003). *Vorhersage von Rechenschwäche in der Grundschule* (2. Auflage). Hamburg: Dr. Kovač.

8 Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie, 30* (4), 201–211. doi:10.1024/1010-0652/a000182

9 Jansen, H., Mannhaupt, G., Marx, H. & Skowronek, H. (2002). *BISC: Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten* (2. Auflage). Göttingen: Hogrefe.

10 Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen.* Göttingen: Hogrefe.

11 Roos, J., Schöler, H. & Treutlein, A. (2007). *Zur prognostischen Validität des Heidelberger Auditiven Screenings in der Einschulungsdiagnostik HASE. Abschlussbericht des Projektes EVER* (S. 33). Heidelberg: Pädagogische Hochschule Heidelberg. Verfügbar unter: http://www.ph-heidelberg.de/wp/schoeler/Datein/Abschlussbericht_EVER-HASE_Feb-2007.pdf

12 Barth, K. & Gomm, B. (2008). Gruppentest zur Früherkennung von Lese- und Rechtschreibschwierigkeiten. Phonologische Bewusstheit bei Kindergartenkindern und Schulanfängern (PB-LRS). In W. Schneider, H. Marx & M. Hasselhorn (Hrsg.), *Diagnostik von Rechtschreibleistungen und -kompetenz* (Band 6, S. 7–43). Göttingen: Hogrefe.

ESM-1: Literatur

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

Tabelle ESM-2

Hauptstudie: Deskriptive Statistiken zur Hauptstudie (mit Erhebungen zur 3. und 8. Klasse) – Anteil der Schülerinnen und Schüler (in Prozent %) auf den Kompetenzstufen und Notenstufen der Prädiktoren und Kriterien separat nach besuchter Schulform in der 8. Klassenstufe

Schülerinnen und Schüler an Schulen mit mehreren Bildungsgängen (SMBG)

Testleistung	KS I	KS II	KS III	KS IV	KS V	
Mathematik						
3. Klasse	27	29	23	13	7	
8. Klasse	49	32	15	3	1	
Deutsch (Lesen)						
3. Klasse	11	20	33	22	15	
8. Klasse	17	36	30	13	4	
Schulnoten	1	2	3	4	5	6
Mathematik						
3. Klasse	5	41	41	12	1	0
8. Klasse	5	26	37	24	7	1
Deutsch						
3. Klasse	5	45	43	7	0	0
8. Klasse	5	32	43	18	3	0

Schülerinnen und Schüler an Gymnasien

Testleistung	KS I	KS II	KS III	KSI V	KS V	
Mathematik						
3. Klasse	4	13	24	25	34	
8. Klasse	5	21	39	21	14	
Deutsch (Lesen)						
3. Klasse	1	4	18	28	50	
8. Klasse	1	8	27	35	29	
Schulnoten	1	2	3	4	5	6
Mathematik						
3. Klasse	32	60	8	0	0	0
8. Klasse	8	30	35	21	5	0
Deutsch						
3. Klasse	31	62	7	0	0	0
8. Klasse	10	44	37	9	1	0

Anmerkungen. KS = Kompetenzstufe.

Tabelle ESM-3

Hauptstudie: Analysen zum Stichprobenausfall hinsichtlich der Prädiktoren – separat nach besuchter Schulform in der 8. Klassenstufe

	Schulen mit mehreren Bildungsgängen (SMBG)					Gymnasien				
	in Analysen einbezogene SuS (N = 6163)		von Analysen ausgeschlossene SuS (N = 475)			in Analysen einbezogene SuS (N = 4776)		von Analysen ausgeschlossene SuS (N = 261)		
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Testleistung Mathematik										
3. Klasse	451	96	454	103	- 0.03	563	109	547	99	0.15
8. Klasse	441	91	442	100	- 0.01	575	92	575	94	0.00
Deutsch (Lesen)										
3. Klasse	514	114	524	125	- 0.08	639	123	632	137	0.05
8. Klasse	495	92	498	104	- 0.03	609	82	609	91	0.00
Schulnoten Mathematik										
3. Klasse	2.62	0.78	2.69	0.82	- 0.09	1.76	0.60	1.80	0.63	- 0.07
8. Klasse	3.05	1.03	3.15	1.07	- 0.10	2.86	1.03	2.89	1.07	- 0.03
Deutsch										
3. Klasse	2.52	0.71	2.65	0.75	- 0.18	1.76	0.57	1.79	0.55	- 0.05
8. Klasse	2.84	0.89	2.85	0.94	- 0.01	2.47	0.82	2.53	0.84	- 0.07
Merkmale in der 3. Klasse										
Alter (in Jahren)	9.58	0.58	9.65	0.59	- 0.12	9.46	0.51	9.42	0.52	0.08
Anteil der Mädchen	0.48	0.50	0.47	0.50	0.02	0.50	0.50	0.48	0.50	0.04
Anteil der Kinder mit Deutsch als Muttersprache	0.98	0.15	0.96	0.20	0.11	0.98	0.14	0.95	0.21	0.17
Anteil der Kinder mit besonderen Lernschwierigkeiten[a] in Mathematik und/oder Deutsch										
3. Klasse	0.13	0.33	0.14	0.34	- 0.03	0.01	0.11	0.02	0.14	- 0.08
8. Klasse	0.06	0.23	0.08	0.27	- 0.08	0.01	0.09	0.02	0.12	- 0.09

Anmerkungen. SuS = Schülerinnen und Schüler, N = Stichprobengröße, M = Mittelwert, SD = Standardabweichung, d = Cohen's d mit gepoolter Standardabweichung. [a] Vorliegen besonderer Lernschwierigkeiten im Schriftspracherwerb, im Verhalten oder im Rechnen.

Tabelle ESM-4

Hauptstudie: Korrelationsmatrix für die Analysestichprobe (für Schülerinnen und Schüler ohne fehlende Werte auf den Prädiktoren) mit Mittelwerten, Standardabweichung und Anteil fehlender Werte über alle imputierten Datensätze

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]
[1]	T_M 3	–	.52	-.48	-.36	.46	.30	-.30	-.20	.03	-.22	.12	-.20	*-.07*
[2]	T_M 8	.53	–	-.39	-.53	.36	.41	-.26	-.35	*-.03*	-.06	.16	-.22	*-.15*
[3]	N_M 3	-.53	-.49	–	.34	-.29	-.20	.39	.22	-.05	.15	-.09	.30	.19
[4]	N_M 8	-.30	-.54	.33	–	-.28	-.33	.28	.49	.05	-.11	*-.07*	.22	.21
[5]	T_D 3	.52	.41	-.38	-.22	–	.35	-.34	-.25	.04	.05	.28	-.20	*-.18*
[6]	T_D 8	.34	.53	-.30	-.33	.43	–	-.27	-.35	*-.02*	.08	.10	-.23	-.24
[7]	N_D 3	-.31	-.30	.45	.26	-.41	-.34	–	.36	-.04	-.22	-.14	.38	.32
[8]	N_D 8	-.22	-.38	.25	.49	-.30	-.41	.39	–	.03	-.37	*-.04*	.21	.27
[9]	Alter 3	-.05	-.16	.08	.13	-.03	-.12	.10	.12	–	-.05	-.09	*.03*	*.02*
[10]	Mäd 3	-.19	-.09	.15	-.07	.06	.11	-.27	-.35	-.05	–	*.03*	*-.05*	-.20
[11]	D-MSp 3	.14	.15	-.10	*-.06*	.11	*.05*	-.15	*-.04*	*.02*	*-.02*	–	-.31	*-.02*
[12]	BL_{DM} 3	-.35	-.33	.44	.22	-.37	-.29	.46	.27	.16	-.09	-.19	–	.37
[13]	BL_{DM} 8	-.21	-.28	.25	.22	-.26	-.25	.28	.29	.12	-.21	*-.08*	0.43	–
	M_{GY}	562	574	1.76	2.86	639	607	1.76	2.48	9.46	0.50	0.98	0.01	0.01
	M_{SMBG}	451	439	2.62	3.05	515	493	2.52	2.84	9.58	0.48	0.98	0.13	0.06
	SD_{GY}	109	93	0.60	1.03	123	83	0.57	0.82	0.51	–	–	–	–
	SD_{SMBG}	96	93	0.78	1.13	114	92	0.71	0.89	0.58	–	–	–	–
	Mis_{GY} (%)	0	8	0	1	0	10	0	4	0	0	0	0	0
	Mis_{SMBG} (%)	0	9	0	1	0	11	0	2	0	0	0	0	0

Anmerkungen. Korrelationswerte für Jugendliche an Gymnasien oberhalb der Hauptdiagonalen. Werte für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) unterhalb der Hauptdiagonalen. In der 1. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung angegeben (bspw. T_M 3 = Erhebung in der 3. Klasse). T_M = Mathematiktestleistung, T_D = Deutschtestleistung, N_M = Mathematiknote, N_D = Deutschnote, Alter = Alter in Jahren, Mäd = Dummyvariable für Geschlecht (0 = Jungen; 1 = Mädchen), D-MSp = Dummyvariable für Deutsch als Muttersprache (0 = andere Sprache als Deutsch; 1 = Deutsch), BL_{DM} = Dummyvariable für das Vorliegen besonderer Lernschwierigkeiten im Schriftspracherwerb, im Verhalten oder im Rechnen in den Fächern Mathematik und/oder Deutsch (0 = nicht vorliegend; 1 = vorliegend). Alle Korrelationen (außer kursiv gesetzte Werte) sind signifikant $p < .05$ (zweiseitig). *M* = Mittelwert, *SD* = Standardabweichung, Mis (%) = Anteil fehlender Werte in Prozent. $_{GY}$ = Schülerinnen und Schüler an Gymnasien. $_{SMBG}$ = Schülerinnen und Schüler an Schulen mit mehreren Bildungsgängen (SMBG).

Details zur Ergänzungsstudie

Im Rahmen der Ergänzungsstudie untersuchten wir auf Grundlage eines Längsschnittdesigns, inwiefern sich die Ergebnisse der Hauptstudie mit Prognosen von der 3. bis 8. Jahrgangsstufe für Prognosen von der 4. bis 5. bzw. 6. Jahrgangsstufe innerhalb der Primarstufe für Mathematik replizieren lassen. Die Rationale zur Begründung der Schwellenwerte für die Klassifikation der beiden Prädiktoren und das Kriterium der Note ist identisch mit der Rationale, die wir im Abschnitt „Schwellen(-werte), Kriterien und Strategien" im Manuskript für die Hauptstudie dargestellt haben. Die Rationale zur Begründung des Schwellenwertes für das Kriterium der Testleistung (Verfehlen des Regelstandards) weicht von der Hauptstudie ab und wird im Manuskript am Ende des Abschnitts „Prognostizierte Kriterien" beschrieben.

Methode

Stichprobe und Prozeduren

Die Ergänzungsstudie basiert auf einem Längsschnittdatensatz mit 430 Kindern (50 % Mädchen, Durchschnittsalter in der 4. Klasse: M = 10.5 Jahre, SD = 0.6 Jahre) an 20 öffentlichen Grundschulen (jeweils eine Klasse) eines Bundeslandes[5]. Kompetenzdaten für das Fach Mathematik lagen für die 4., 5. und 6. Klassenstufe vor und wurden in den Schuljahren 2008/09 bis 2010/2011 erhoben. Die Datenerhebungen fanden jeweils gegen Ende des zweiten Schulhalbjahres in den Monaten Mai und Juli statt (für Details siehe Abschlussbericht zur Studie Autor, 2014). In die Analysen bezogen wir nur Kinder ein, für die in der 4. Klasse sowohl Testergebnisse als auch Noten für das Fach Mathematik vorlagen. Es zeigten sich (bemessen an Cohen's d) maximal kleinere, systematische Unterschiede zugunsten der Kinder der Analysestichprobe (bessere Testleistung und Schulnoten) im Vergleich zu jenen, die nicht in die Analysen aufgenommen wurden (Tabelle ESM-5.1). Aufgrund der positiv selektierten Stichprobe im Rahmen der Ergänzungsstudie könnte möglicherweise eine systematische Unterschätzung der klassifikatorischen Indizes, v. a. des positiv prädiktiven Wertes, vorliegen (Tröster, 2009). Weitere Details zur Stichprobe in Tabelle ESM-5.2 und Tabelle ESM-5.3.

[5] Um keine neuartigen Vergleiche zu ermöglichen, wird das Bundesland nicht benannt.

Tabelle ESM-5.1

Analysen zum Stichprobenausfall hinsichtlich der Prädiktoren

	in Analysen einbezogene SuS (N = 430)		von Analysen ausgeschlossene SuS (N = 68)		
	M	*SD*	*M*	*SD*	*d*
Testleistung: Mathematik					
4. Klasse	531	107	493	102	0.36
5. Klasse	577	117	543	105	0.31
6. Klasse	601	120	573	116	0.24
Schulnote: Mathematik					
4. Klasse	2.22	0.91	2.54	0.95	- 0.34
5. Klasse	2.41	0.96	2.69	0.98	- 0.29
6. Klasse	2.37	1.03	2.60	0.88	- 0.24
Merkmale in der 4. Klasse					
Alter (in Jahren)	10.48	0.57	10.48	0.54	0.00
Anteil der Mädchen	0.50	0.50	0.49	0.50	0.02

Anmerkungen. SuS = Schülerinnen und Schüler, N = Stichprobengröße, M = Mittelwert, SD = Standardabweichung, d = Cohen's d mit gepoolter Standardabweichung.

Tabelle ESM-5.2

Deskriptive Statistiken zur Ergänzungsstudie (mit Erhebungen zur 4., 5. und 6. Klasse) – Anteil der Schülerinnen und Schüler (in Prozent %) auf die Kompetenzstufen und Notenstufen der Prädiktoren und Kriterien

Testleistung	KS I	KS II	KS III	KS IV	KS V	
Mathematik						
4. Klasse	8	19	20	26	28	
5. Klasse	5	12	20	22	42	
6. Klasse	2	12	15	21	50	
Schulnoten	1	2	3	4	5	6
Mathematik						
4. Klasse	21	46	24	7	2	0
5. Klasse	16	43	29	10	3	0
6. Klasse	21	38	27	11	3	0

Anmerkungen. KS = Kompetenzstufe.

Messinstrumente

In der Ergänzungsstudie wurde die mathematische Kompetenz für vier Leitideen erfasst (21/29/41 Testitems in 4./5./6. Klassenstufe): Zahlen und Operationen, Muster und Strukturen, Raum und Form sowie Größen und Messen. Alle Testitems stammten aus einem Aufgabenpool, der vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) zur Überprüfung der Bildungsstandards in der Primarstufe entwickelt wurde. Die Testitems der 4. Klasse sind auch als kommerzieller Test erhältlich (Granzer et al., 2008a). Weitere Informationen zur Testkonzeption und zur Testgüte finden sich unter anderem in der Handreichung zu den Testheften (Granzer et al., 2008b; Köller, Eβel-Ullmann & Paasch, 2012). Die vorgeschriebene Bearbeitungsdauer pro Test betrug maximal 40 Minuten. Die Reliabilität (interne Konsistenz, KR-20) lag bei r_{tt} = .80, .88 und .91 in der 4., 5. und 6. Klassenstufe. Diese Werte sind vergleichbar mit bisherigen Angaben zur Reliabilität (Köller et al., 2012). Die Durchführung der Tests sowie die Auswertung und Ergebniseingabe erfolgte durch geschultes Personal. In beiden Studien gaben Lehrkräfte für die Schülerinnen und Schüler zu allen untersuchten Erhebungszeitpunkten die Halbjahresnote im Fach Mathematik auf der Schulnotenskala von 1 (*sehr gut*) bis 6 (*ungenügend*) an (s. Interkorrelationstabelle ESM-5.3).

Datenanalyse

Skalierung der bildungsstandardbasierten Testergebnisse

Im Rahmen der Ergänzungsstudie erfolgte die Skalierung der bildungsstandardbasierten Kompetenztests für Mathematik auf Basis eines eindimensionalen Rasch-Modells (Rasch, 1980) für dichotome Daten. Hierzu wurden die Itemparameter aus den Pilotierungsstudien des IQB herangezogen. Die Personenparameter wurden für die nachfolgenden Analysen unter Nutzung der Software R (Paket: sirt) auf die Bildungsstandardmetrik transformiert, um die erreichten Kompetenzstufen zu bestimmen.

Für die längsschnittlichen Analysen im Rahmen der Ergänzungsstudie wurde ein Ankeritemdesign zur Messung der mathematischen Kompetenz verwendet (s. Abbildung ESM-5.1). In den Analysen wurden nur Items berücksichtigt, die ein Mindestmaß an Messäquivalenz über die Zeit aufwiesen. Eine ausführliche Darstellung der Itemanalyse, des Linking und der Parameterschätzung wird ausführlich in Autor (2014) berichtet.

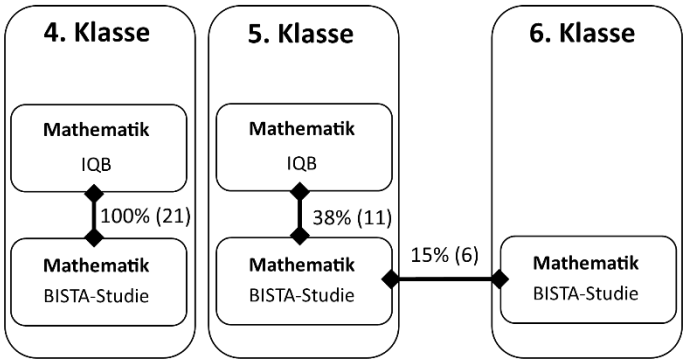


Abbildung ESM-5.1. Veranschaulichung der Messinvarianz „Verlinkung“ der bildungsstandardbasierten Kompetenztests in Mathematik der Ergänzungsstudie (= BISTA-Studie) mit Itemparametern aus der IQB-Pilotierungsstichprobe (= IQB). Die Prozentangaben beziehen sich auf den Anteil der „verlinkten“ Aufgaben im bildungsstandardbasierten Test, für den Itemparameter aus den Referenzstudien vorlagen. In Klammern steht die Anzahl an sogenannten „Linking“-Items. Lesebeispiel der Abbildung: Für 100 % der Items (also 21 Items) für den bildungsstandardbasierten Mathematiktest der Ergänzungsstudie in der 4. Klasse lagen Itemparameter aus der Pilotierungsstudie des IQB vor. Von Autor, 2017, Copyright 2017 bei Hogrefe. Veränderte Wiedergabe.

Umgang mit fehlenden Werten

Der Anteil fehlender Werte pro Variable lag zwischen 17 und 23 % (Tabelle ESM-5.3). Analog zur Hauptstudie nutzten wir das multiple Imputationsverfahren MICE (Buuren & Groothuis-Oudshoorn, 2011), um jeweils 15 vollständige Datensätze zu erzeugen. Bei der Imputation der fehlenden Werte berücksichtigten wir das Skalenniveau der einbezogenen Variablen sowie die hierarchische Datenstruktur (Schulebene). Um die Qualität der

imputierten Daten zu verbessern, nutzten wir in der Ergänzungsstudie neben den untersuchten Prädiktoren und Kriterien folgende Merkmale als Hilfsvariablen (Collins, Schafer & Kam, 2001): Alter, Geschlecht, Mathematiktestleistung in der 2. und 3. Klasse, Deutschtestleistung (Lesen) in bildungsstandardbasierten Tests in der 4., 5. und 6. Klasse, Deutschnote in der 4., 5. und 6. Klasse, Testergebnis zu Figurenanalogien (N2), Höchster ISEI-Wert der Eltern; monatliches Haushaltsnettoeinkommen, Buchbesitz, schulischer Abschluss der Eltern; beruflicher Abschluss der Eltern, kulturelle Aktivitäten, Summe der Wohlstandsgüter, Gymnasialempfehlung (für Details siehe Autor, 2017).

Tabelle ESM-5.3

Korrelationsmatrix für die Analysestichprobe (für Schülerinnen und Schüler ohne fehlende Werte auf den Prädiktoren) mit Mittelwerten, Standardabweichung und Anteil fehlender Werte über alle imputierten Datensätze

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
[1]	T_M 4	–	.64	.61	-.59	-.55	-.57	-.12	*-.06*
[2]	T_M 5		–	.75	-.63	-.62	-.67	-.19	-.20
[3]	T_M 6			–	-.64	-.62	-.67	-.17	-.15
[4]	N_M 4				–	.72	.71	.20	.10
[5]	N_M 5					–	.74	.18	*.05*
[6]	N_M 6						–	.28	*.07*
[7]	Alter 4							–	*-.08*
[8]	Mäd								–
	M	531	580	603	2.22	2.39	2.39	10.47	0.50
	SD	107	118	124	0.91	0.98	1.07	0.58	–
	Mis (%)	0	17	23	0	17	19	17	0

Anmerkungen. In der 2. Spalte nach der Variablenabkürzung ist die Klassenstufe der Erhebung angegeben (bspw. T_M 4 = Erhebung in der 4. Klasse). T_M = Mathematiktestleistung, N_M = Mathematiknote, Alter = Alter in Jahren, Mäd = Dummyvariable für Geschlecht (0 = Jungen; 1 = Mädchen). Alle Korrelationen (außer kursiv gesetzte Werte) sind signifikant $p < .05$ (zweiseitig). *M* = Mittelwert, *SD* = Standardabweichung, Mis (%) = Anteil fehlender Werte in Prozent.

ESM-5: Literatur

Autor[6], 2014

Autor6, 2017

Buuren, S. van & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal Of Statistical Software*, *45* (3), 1–67.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6* (4), 330–351. doi:10.1037/1082-989X.6.4.330

Fuchs, G. & Brunner, M. (2014, September). Dissertationsvorhaben: Psychometrische Datenqualität bildungsstandardbasierter Testverfahren. Vortrag gehalten auf der 1. BIEN-Nachwuchstagung, Berlin.

Granzer, D., Köller, O., Reiss, K., Robitzsch, A., Walther, G. & Winkelmann, H. (2008a). *Bildungsstandards. Kompetenzen überprüfen. Grundschule. Klasse 3/4 – Heft 2*. Berlin: Cornelsen Verlag.

Granzer, D., Köller, O., Reiss, K., Robitzsch, A., Walther, G. & Winkelmann, H. (2008b). *Bildungsstandards. Kompetenzen überprüfen. Grundschule. Klasse 3/4 – Handreichung*. Berlin: Cornelsen Verlag.

Köller, O., Eβel-Ullmann, G. & Paasch, D. (2012). Validierung eines Instruments zur Erfassung Standard-basierter mathematischer Kompetenzen in der Grundschule. *Psychologie in Erziehung und Unterricht*, *59* (3), 177–190.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Edition.). Chicago: University of Chicago Press.

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

[6] Die allgemeine Literaturangabe dient der Anonymisierung der Daten.

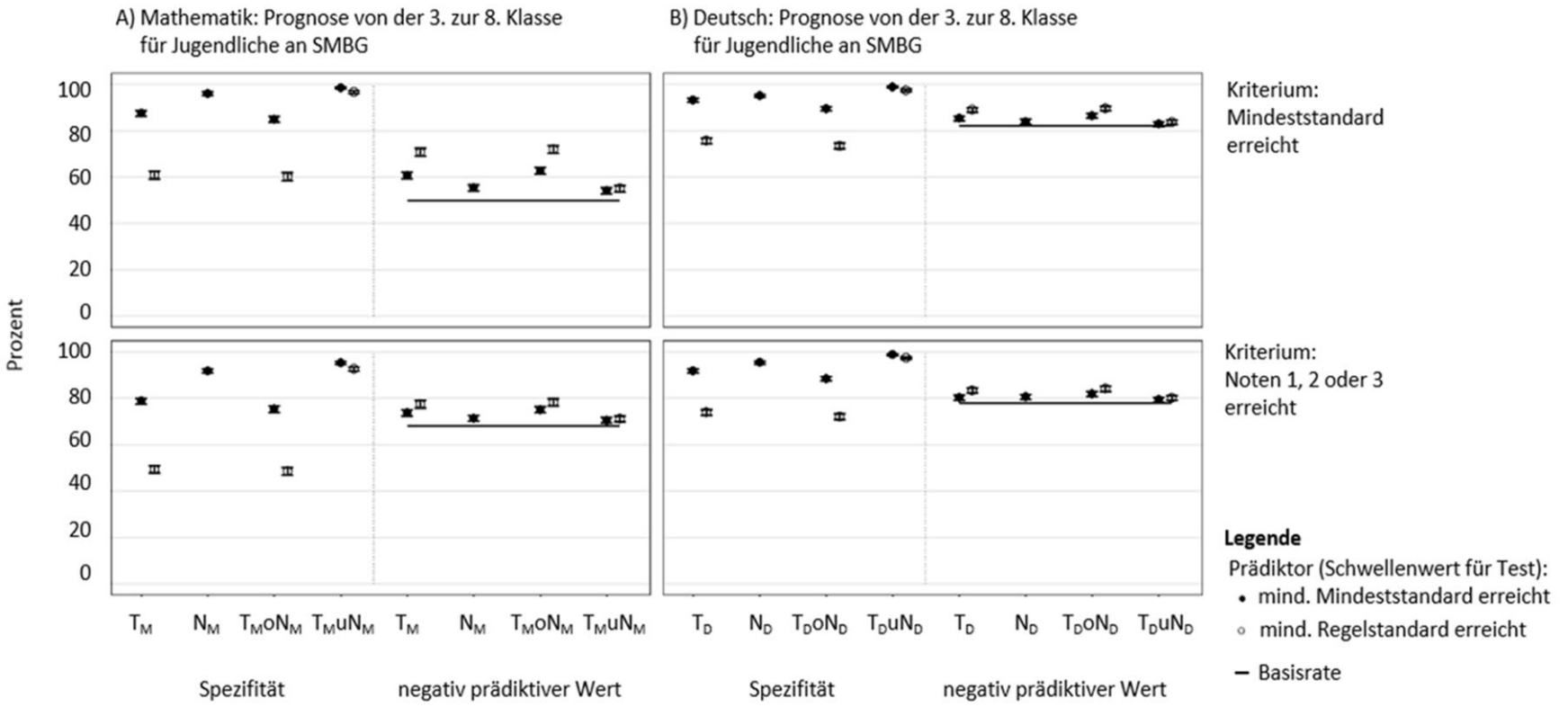


Abbildung ESM-6. Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach (A) Mathematik und (B) Deutsch: Spezifität (*Spe*) und negativ prädiktive Werte (*npW*, in Prozent) für das Erreichen des Mindeststandards (= mind. Mindeststandard erreicht) und das Erreichen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6 verfehlt) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Prädiktorkombinationen (Spezifikation s. Text) und (2) der Schwelle für die Kompetenztests (Erreichen der Mindest- bzw. Regelstandards). T_M bzw. T_D = Testergebnis in Mathematik bzw. Testergebnis in Deutsch, N_M bzw. N_D = Schulnote in Mathematik bzw. Schulnote in Deutsch, $T_M oN_M$ bzw. $T_D oN_D$ = Mathematiktestergebnis oder Mathematiknote bzw. Deutschtestergebnis oder Deutschnote, $T_M uN_M$ bzw. $T_D uN_D$ = Mathematiktestergebnis und Mathematiknote bzw. Deutschtestergebnis und Deutschnote. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Jugendlichen, die in der 8. Klasse ein bestimmtes mathematik- bzw. deutschbezogenes Bildungsergebnis nicht verfehlten bzw. erreichten).

Tabelle ESM-7

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „nicht gefährdeter“ Kinder durch Erreichen des Mindeststandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	Spezifität (in %)		negativ prädiktiver Wert (in %)		Spezifität (in %)		negativ prädiktiver Wert (in %)	
Mathematik/Deutsch	Mathematik/Deutsch	Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Mindeststandard erreicht (BR = 50 %/82 %)	T_M/T_D (SR = 73 %/89 %)	87	[86;89]	61	[59;62]	93	[92;94]	85	[84;86]
	N_M/N_D (SR = 87 %/93 %)	96	[95;97]	55	[54;57]	95	[94;96]	84	[83;85]
	T_MoN_M/T_DoN_D (SR = 68 %/85 %)	85	[84;86]	63	[61;64]	90	[89;90]	86	[86;87]
	T_MuN_M/T_DuN_D (SR = 92 %/98 %)	99	[98;99]	54	[53;55]	99	[98;99]	83	[82;84]
Noten 1, 2 oder 3 erreicht (BR = 68 %/78 %)	T_M/T_D (SR = 73 %/89 %)	79	[77;80]	74	[72;75]	92	[91;92]	80	[79;81]
	N_M/N_D (SR = 87 %/93 %)	92	[91;93]	71	[70;73]	95	[95;96]	81	[80;82]
	T_MoN_M/T_DoN_D (SR = 68 %/85 %)	75	[74;77]	75	[74;76]	88	[87;89]	82	[81;83]
	T_MuN_M/T_DuN_D (SR = 92 %/98 %)	95	[95;96]	70	[69;72]	99	[98;99]	79	[78;80]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, $T_M/_{TD}$: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, T_MoN_M/T_DoN_D: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, T_MuN_M/T_DuN_D: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-8

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach (A) Mathematik bzw. (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „nicht gefährdeter“ Kinder durch Erreichen des Regelstandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse Mathematik/Deutsch	Prädiktor(en) 3. Klasse Mathematik/Deutsch	Spezifität (in %)		negativ prädiktiver Wert (in %)		Spezifität (in %)		negativ prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Mindeststandard erreicht (BR = 50 %/82 %)	T_M/T_D (SR = 43 %/70 %)	61	[59;63]	71	[69;73]	76	[75;77]	89	[88;90]
	N_M/N_D (SR = 87 %/93 %)	96	[95;97]	55	[54;57]	95	[94;96]	84	[83;85]
	T_MoN_M/T_DoN_D (SR = 42 %/67 %)	60	[58;62]	72	[70;74]	73	[72;75]	90	[89;90]
	T_MuN_M/T_DuN_D (SR = 89 %/95 %)	97	[96;97]	55	[54;56]	97	[97;98]	84	[83;85]
Noten 1, 2 oder 3 erreicht (BR = 68 %/78 %)	T_M/T_D (SR = 43 %/70 %)	49	[48;51]	77	[76;79]	74	[73;75]	83	[82;84]
	N_M/N_D (SR = 87 %/93 %)	92	[91;93]	71	[70;73]	95	[95;96]	81	[80;82]
	T_MoN_M/T_DoN_D (SR = 42 %/67 %)	48	[47;50]	78	[77;80]	72	[71;73]	84	[83;85]
	T_MuN_M/T_DuN_D (SR = 89 %/95 %)	93	[92;93]	71	[70;72]	97	[97;98]	80	[79;81]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M/T_D: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, T_MoN_M/T_DoN_D: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, T_MuN_M/T_DuN_D: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

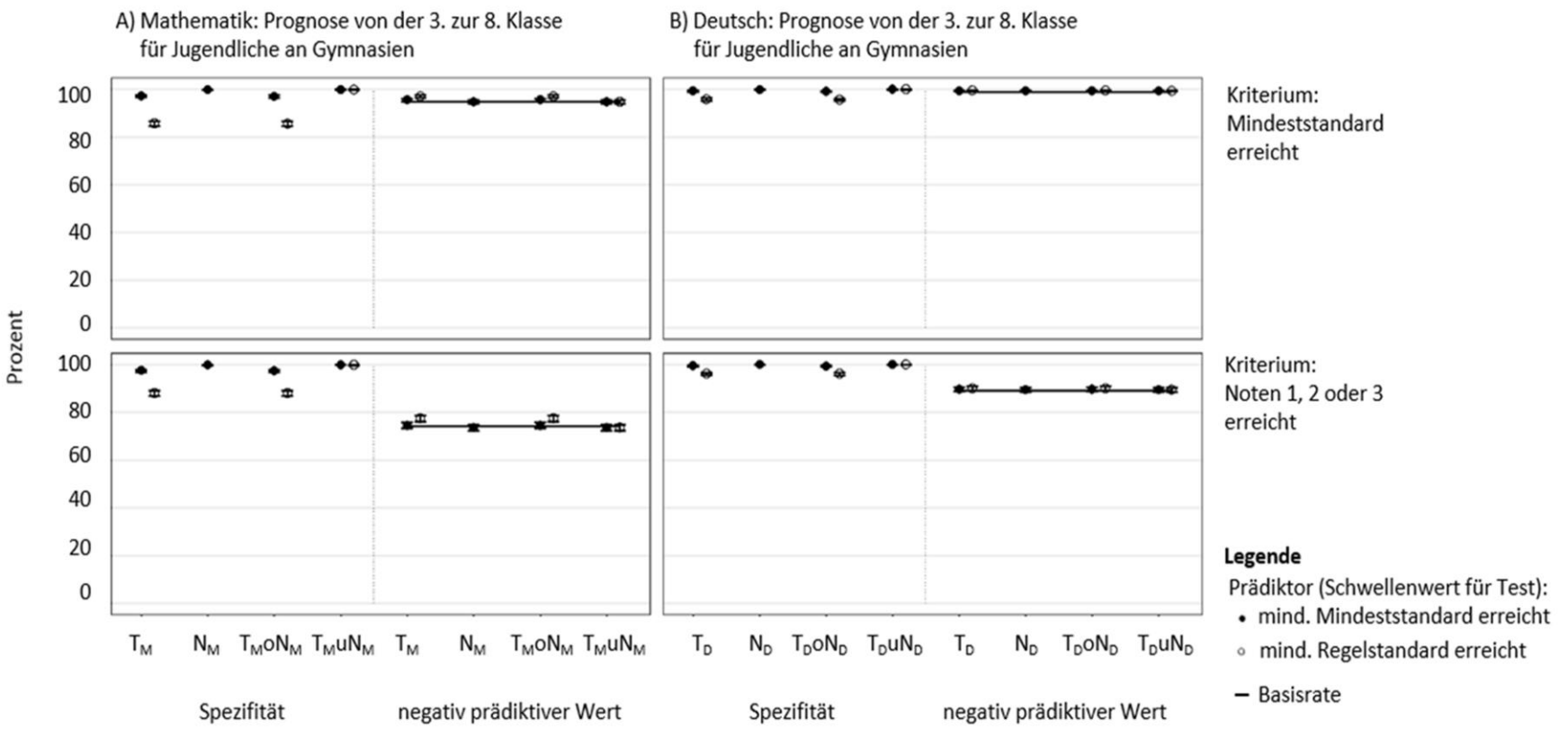


Abbildung ESM-9. Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach (A) Mathematik und (B) Deutsch: Spezifität (*Spe*) und negativ prädiktive Werte (*npW*, in Prozent) für das Erreichen des Mindeststandards (= mind. Mindeststandard erreicht) und das Erreichen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6 verfehlt) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Prädiktorkombinationen (Spezifikation s. Text) und (2) der Schwelle für die Kompetenztests (Erreichen der Mindest- bzw. Regelstandards). T_M bzw. T_D = Testergebnis in Mathematik bzw. Testergebnis in Deutsch, N_M bzw. N_D = Schulnote in Mathematik bzw. Schulnote in Deutsch, T_MoN_M bzw. T_DoN_D = Mathematiktestergebnis oder Mathematiknote bzw. Deutschtestergebnis oder Deutschnote, T_MuN_M bzw. T_DuN_D = Mathematiktestergebnis und Mathematiknote bzw. Deutschtestergebnis und Deutschnote. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Jugendlichen, die in der 8. Klasse ein bestimmtes mathematik- bzw. deutschbezogenes Bildungsergebnis nicht verfehlten bzw. erreichten).

Tabelle ESM-10

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „nicht gefährdeter“ Kinder durch Erreichen des Mindeststandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	Spezifität (in %)		negativ prädiktiver Wert (in %)		Spezifität (in %)		negativ prädiktiver Wert (in %)	
Mathematik/Deutsch	Mathematik/Deutsch	Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Mindeststandard erreicht (BR = 95 %/99 %)	T_M/T_D (SR = 96 %/99 %)	97	[97;98]	96	[95;96]	99	[99;99]	99	[99;100]
	N_M/N_D (SR = 100 %/100 %)	100	[100;100]	95	[94;95]	100	[100;100]	99	[99;100]
	T_MoN_M/T_DoN_D (SR = 96 %/99 %)	97	[97;98]	96	[95;96]	99	[99;99]	99	[99;100]
	T_MuN_M/T_DuN_D (SR = 100 %/100 %)	100	[100;100]	95	[94;95]	100	[100;100]	99	[99;100]
Noten 1, 2 oder 3 erreicht (BR = 74 %/89 %)	T_M/T_D (SR = 96 %/99 %)	97	[97;98]	74	[73;76]	99	[99;100]	90	[89;90]
	N_M/N_D (SR = 100 %/100 %)	100	[100;100]	74	[72;75]	100	[100;100]	89	[89;90]
	T_MoN_M/T_DoN_D (SR = 96 %/99 %)	97	[97;98]	74	[73;76]	99	[99;100]	90	[89;90]
	T_MuN_M/T_DuN_D (SR = 100 %/100 %)	100	[100;100]	74	[72;75]	100	[100;100]	89	[88;90]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M/T_D: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, T_MoN_M/T_DoN_D: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, T_MuN_M/T_DuN_D: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-11

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „nicht gefährdeter“ Kinder durch Erreichen des Regelstandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	Spezifität (in %)		negativ prädiktiver Wert (in %)		Spezifität (in %)		negativ prädiktiver Wert (in %)	
Mathematik/Deutsch	Mathematik/Deutsch	Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Mindeststandard erreicht (BR = 95 %/99 %)	T_M/T_D (SR = 84 %/96 %)	86	[85;87]	97	[97;98]	96	[95;96]	99	[99;100]
	N_M/N_D (SR = 100 %/100 %)	100	[100;100]	95	[94;95]	100	[100;100]	99	[99;100]
	T_MoN_M/T_DoN_D (SR = 84 %/96 %)	86	[85;87]	97	[97;98]	96	[95;96]	99	[99;100]
	T_MuN_M/T_DuN_D (SR = 100 %/100 %)	100	[100;100]	95	[94;95]	100	[100;100]	99	[99;100]
Noten 1, 2 oder 3 erreicht (BR = 74 %/89 %)	T_M/T_D (SR = 84 %/96 %)	88	[87;89]	77	[76;79]	96	[96;97]	90	[89;91]
	N_M/N_D (SR = 100 %/100 %)	100	[100;100]	74	[72;75]	100	[100;100]	89	[89;90]
	T_MoN_M/T_DoN_D (SR = 84 %/96 %)	88	[87;89]	77	[76;79]	96	[96;97]	90	[89;91]
	T_MuN_M/T_DuN_D (SR = 100 %/100 %)	100	[100;100]	74	[72;75]	100	[100;100]	89	[88;90]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M/T_D: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, T_MoN_M/T_DoN_D: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, T_MuN_M/T_DuN_D: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

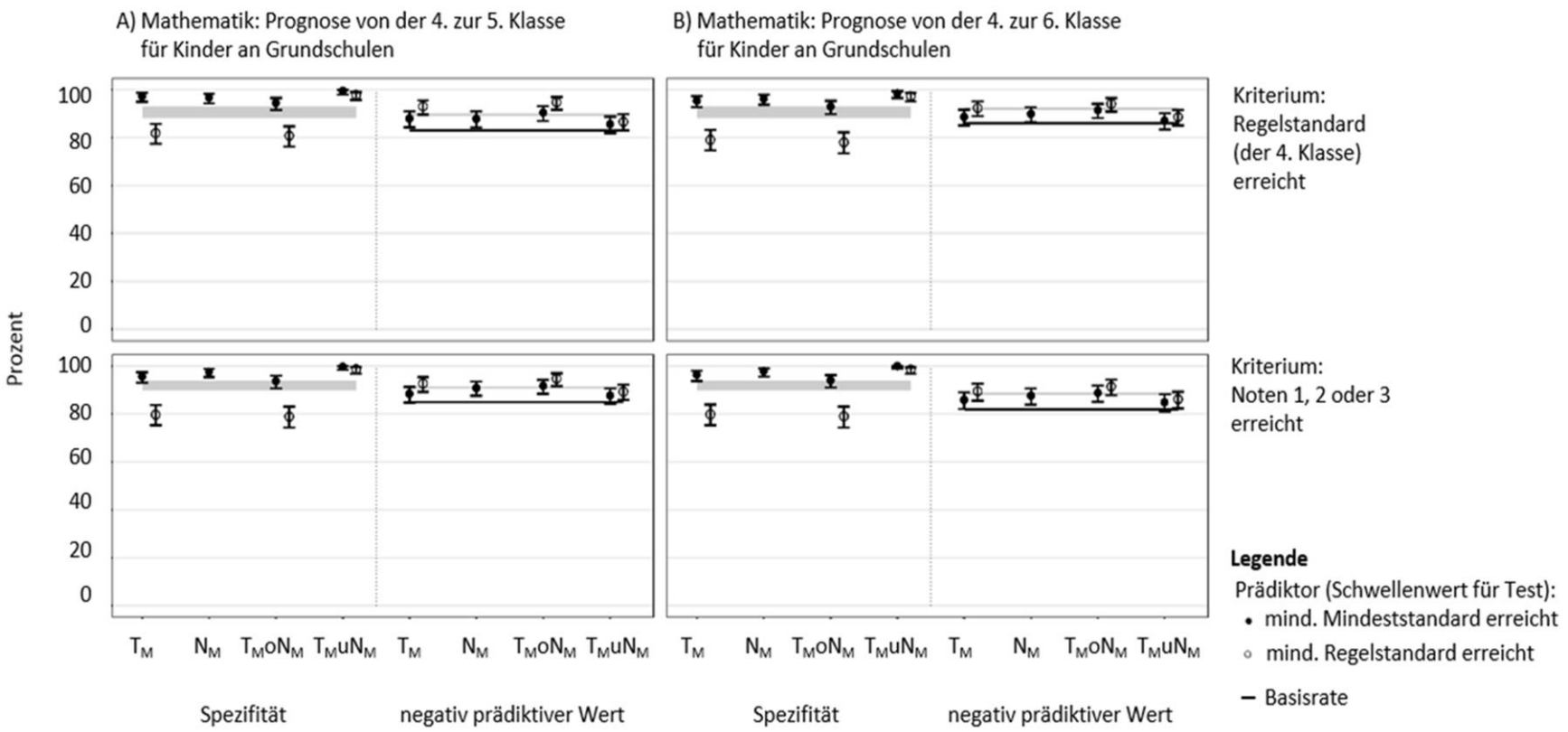


Abbildung ESM-12. Ergebnisse der Ergänzungsstudie für Kinder an Grundschulen im Fach Mathematik für (A) Prognosen von der 4. zur 5. Klasse und (B) Prognosen von der 4. zur 6. Klasse: Spezifität (*Spe*) und negativ prädiktive Werte (*npW*, in Prozent) für das Erreichen des Regelstandards (= mind. Regelstandard erreicht, definiert für die 4. Jahrgangsstufe) und das Erreichen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6 verfehlt) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Prädiktorkombinationen (Spezifikation s. Text) und (2) der Schwelle für die Kompetenztests (Erreichen der Mindest- bzw. Regelstandards). T_M = Testergebnis in Mathematik, N_M = Schulnote in Mathematik, T_MoN_M = Mathematiktestergebnis oder Mathematiknote, T_MuN_M = Mathematiktestergebnis und Mathematiknote. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Kinder, die in der 5. bzw. 6. Klasse ein bestimmtes mathematikbezogenes Bildungsergebnis nicht verfehlten bzw. erreichten). Die grau unterlegten Werte markieren den Referenzbereich, der für deutschsprachige standardisierte Schulleistungstests ermittelt wurde (für Details s. ESM-36).

Tabelle ESM-13

Ergebnisse für die Ergänzungsstudie für Kinder an Grundschulen im Fach Mathematik für (A) Prognosen von der 4. zur 5. Klasse und (B) Prognosen von der 4. zur 6. Klasse: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „nicht gefährdeter“ Kinder durch Erreichen des Mindeststandards in 4. Klasse

		Prognose von der 4. Klasse zur 5. Klasse				Prognose von der 4. Klasse zur 6. Klasse			
Kriterium 5. Klasse/6. Klasse	Prädiktor(en) 4. Klasse	Spezifität (in %)		negativ prädiktiver Wert (in %)		Spezifität (in %)		negativ prädiktiver Wert (in %)	
Mathematik	Mathematik	Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Regelstandard (der 4. Klasse) erreicht (BR = 87 %/89 %)	T_M (SR = 92 %/92 %)	97	[95;99]	88	[84;91]	95	[93;97]	89	[85;92]
	N_M (SR = 92 %/92 %)	97	[94;98]	88	[84;91]	96	[94;98]	90	[86;93]
	$T_M o N_M$ (SR = 87 %/87 %)	94	[92;97]	90	[87;93]	93	[90;95]	91	[88;94]
	$T_M u N_M$ (SR = 97 %/97 %)	99	[98;100]	86	[82;89]	98	[96;99]	87	[83;90]
Noten 1, 2 oder 3 erreicht (BR = 88 %/86 %)	T_M (SR = 92 %/92 %)	96	[93;97]	88	[85;91]	96	[94;98]	86	[82;89]
	N_M (SR = 92 %/92 %)	98	[95;99]	91	[88;94]	98	[96;99]	88	[84;91]
	$T_M o N_M$ (SR = 87 %/87 %)	94	[91;96]	92	[88;94]	94	[91;96]	89	[85;92]
	$T_M u N_M$ (SR = 97 %/97 %)	100	[98;100]	88	[84;91]	100	[99;100]	85	[81;88]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-14

Ergebnisse für die Ergänzungsstudie für Kinder an Grundschulen im Fach Mathematik für (A) Prognosen von der 4. zur 5. Klasse und (B) Prognosen von der 4. zur 6. Klasse: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „nicht gefährdeter“ Kinder durch Erreichen des Regelstandards in 4. Klasse

		Prognose von der 4. Klasse zur 5. Klasse				Prognose von der 4. Klasse zur 6. Klasse			
Kriterium 5. Klasse/6. Klasse Mathematik	Prädiktor(en) 4. Klasse Mathematik	Spezifität (in %)		negativ prädiktiver Wert (in %)		Spezifität (in %)		negativ prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Regelstandard (der 4. Klasse) erreicht (BR = 87 %/89 %)	T_M (SR = 73 %/73 %)	82	[77;86]	93	[90;96]	79	[75;83]	92	[89;95]
	N_M (SR = 92 %/92 %)	97	[94;98]	88	[84;91]	96	[94;98]	90	[86;93]
	$T_M o N_M$ (SR = 71 %/71 %)	81	[76;85]	95	[92;97]	78	[73;82]	94	[91;96]
	$T_M u N_M$ (SR = 94 %/94 %)	98	[96;99]	87	[83;90]	97	[95;99]	89	[85;92]
Noten 1, 2 oder 3 erreicht (BR = 88 %/86 %)	T_M (SR = 73 %/73 %)	80	[75;84]	93	[89;95]	80	[75;84]	90	[86;93]
	N_M (SR = 92 %/92 %)	98	[95;99]	91	[88;94]	98	[96;99]	88	[84;91]
	$T_M o N_M$ (SR = 71 %/71 %)	79	[74;83]	95	[92;97]	79	[74;83]	91	[88;94]
	$T_M u N_M$ (SR = 94 %/94 %)	99	[97;100]	89	[86;92]	99	[97;100]	86	[82;89]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-15 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 50 %)	T_M (SR = 27 %)	1299	390	1759	2715	21	6	29	44	42	[41;44]	87	[86;89]	77	[75;79]	61	[59;62]
	N_M (SR = 13 %)	651	126	2407	2979	11	2	39	48	21	[20;23]	96	[95;97]	84	[81;86]	55	[54;57]
	T_MoN_M (SR = 32 %)	1489	470	1569	2635	24	8	25	43	49	[47;50]	85	[84;86]	76	[74;78]	63	[61;64]
	T_MuN_M (SR = 8 %)	461	46	2597	3059	7	1	42	50	15	[14;16]	99	[98;99]	91	[88;93]	54	[53;55]
Verfehlen der Noten 1, 2 oder 3 (BR = 32 %)	T_M (SR = 27 %)	800	889	1178	3296	13	14	19	53	40	[38;43]	79	[77;80]	47	[45;50]	74	[72;75]
	N_M (SR = 13 %)	433	344	1545	3841	7	6	25	62	22	[20;24]	92	[91;93]	56	[52;59]	71	[70;73]
	T_MoN_M (SR = 32 %)	926	1033	1051	3153	15	17	17	51	47	[45;49]	75	[74;77]	47	[45;50]	75	[74;76]
	T_MuN_M (SR = 8 %)	306	201	1671	3985	5	3	27	65	15	[14;17]	95	[95;96]	60	[56;65]	70	[69;72]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, T_MoN_M: Testergebnis in Mathematik oder Schulnote in Mathematik, T_MuN_M: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-15 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 50 %)	T_M (SR = 27 %)	3,38	[3,06;3,74]	0,66	[0,64;0,68]	5,14	[4,52;5,84]	0,65	[0,64;0,66]	0,3	[0,27;0,33]	54
	N_M (SR = 13 %)	5,25	[4,36;6,31]	0,82	[0,80;0,84]	6,39	[5,24;7,80]	0,59	[0,58;0,60]	0,17	[0,15;0,19]	68
	$T_M o N_M$ (SR = 32 %)	3,22	[2,94;3,52]	0,6	[0,58;0,63]	5,32	[4,71;6,01]	0,67	[0,66;0,68]	0,34	[0,30;0,37]	52
	$T_M u N_M$ (SR = 8 %)	10,2	[7,55;13,72]	0,86	[0,85;0,88]	11,8	[8,68;16,06]	0,57	[0,56;0,58]	0,14	[0,12;0,15]	82
Verfehlen der Noten 1, 2 oder 3 (BR = 32 %)	T_M (SR = 27 %)	1,9	[1,76;2,06]	0,76	[0,73;0,79]	2,52	[2,24;2,83]	0,66	[0,65;0,68]	0,19	[0,16;0,23]	22
	N_M (SR = 13 %)	2,66	[2,34;3,04]	0,85	[0,83;0,87]	3,13	[2,68;3,65]	0,69	[0,68;0,70]	0,14	[0,11;0,16]	35
	$T_M o N_M$ (SR = 32 %)	1,9	[1,77;2,04]	0,71	[0,67;0,74]	2,69	[2,40;3,01]	0,66	[0,65;0,67]	0,22	[0,19;0,26]	22
	$T_M u N_M$ (SR = 8 %)	3,22	[2,72;3,82]	0,89	[0,87;0,91]	3,63	[3,01;4,38]	0,7	[0,68;0,71]	0,11	[0,08;0,13]	42

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-16 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 18 %)	T_D (SR = 11 %)	306	343	810	4704	5	6	13	76	27	[25;30]	93	[92;94]	47	[43;51]	85	[84;86]
	N_D (SR = 7 %)	189	246	926	4802	3	4	15	78	17	[15;19]	95	[94;96]	43	[39;48]	84	[83;85]
	T_DoN_D (SR = 15 %)	410	529	706	4518	7	9	11	73	37	[34;40]	90	[89;90]	44	[40;47]	86	[86;87]
	T_DuN_D (SR = 2 %)	85	60	1030	4988	1	1	17	81	8	[6;9]	99	[98;99]	59	[50;67]	83	[82;84]
Verfehlen der Noten 1, 2 oder 3 (BR = 22 %)	T_D (SR = 11 %)	245	404	1084	4430	4	7	18	72	18	[16;21]	92	[91;92]	38	[34;42]	80	[79;81]
	N_D (SR = 7 %)	214	221	1114	4614	3	4	18	75	16	[14;18]	95	[95;96]	49	[44;54]	81	[80;82]
	T_DoN_D (SR = 15 %)	377	562	952	4272	6	9	15	69	28	[26;31]	88	[87;89]	40	[37;43]	82	[81;83]
	T_DuN_D (SR = 2 %)	82	63	1246	4772	1	1	20	77	6	[5;8]	99	[98;99]	57	[48;65]	79	[78;80]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-16 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 18 %)	T_D (SR = 11 %)	4,03	[3,51;4,64]	0,78	[0,75;0,81]	5,18	[4,37;6,15]	0,81	[0,80;0,82]	0,21	[0,17;0,24]	35
	N_D (SR = 7 %)	3,48	[2,91;4,16]	0,87	[0,85;0,90]	3,98	[3,25;4,88]	0,81	[0,80;0,82]	0,12	[0,09;0,15]	31
	T_DoN_D (SR = 15 %)	3,51	[3,14;3,92]	0,71	[0,68;0,74]	4,96	[4,26;5,77]	0,8	[0,79;0,81]	0,26	[0,23;0,30]	31
	T_DuN_D (SR = 2 %)	6,41	[4,64;8,87]	0,93	[0,92;0,95]	6,86	[4,90;9,61]	0,82	[0,81;0,83]	0,06	[0,05;0,08]	49
Verfehlen der Noten 1, 2 oder 3 (BR = 22 %)	T_D (SR = 11 %)	2,21	[1,90;2,55]	0,89	[0,87;0,91]	2,48	[2,09;2,94]	0,76	[0,75;0,77]	0,1	[0,07;0,13]	21
	N_D (SR = 7 %)	3,53	[2,95;4,21]	0,88	[0,86;0,90]	4,01	[3,29;4,89]	0,78	[0,77;0,79]	0,12	[0,09;0,14]	35
	T_DoN_D (SR = 15 %)	2,44	[2,17;2,74]	0,81	[0,78;0,84]	3,01	[2,60;3,49]	0,75	[0,74;0,77]	0,17	[0,13;0,20]	24
	T_DuN_D (SR = 2 %)	4,74	[3,43;6,54]	0,95	[0,94;0,96]	4,98	[3,57;6,96]	0,79	[0,78;0,80]	0,05	[0,03;0,07]	45

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-17 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 50 %)	T_M (SR = 57 %)	2281	1216	778	1888	37	20	13	31	75	[73;76]	61	[59;63]	65	[64;67]	71	[69;73]
	N_M (SR = 13 %)	651	126	2407	2979	11	2	39	48	21	[20;23]	96	[95;97]	84	[81;86]	55	[54;57]
	$T_M o N_M$ (SR = 58 %)	2329	1237	729	1868	38	20	12	30	76	[75;78]	60	[58;62]	65	[64;67]	72	[70;74]
	$T_M u N_M$ (SR = 11 %)	603	105	2455	3000	10	2	40	49	20	[18;21]	97	[96;97]	85	[82;88]	55	[54;56]
Verfehlen der Noten 1, 2 oder 3 (BR = 32 %)	T_M (SR = 57 %)	1374	2123	603	2063	22	34	10	33	69	[67;72]	49	[48;51]	39	[38;41]	77	[76;79]
	N_M (SR = 13 %)	433	344	1545	3841	7	6	25	62	22	[20;24]	92	[91;93]	56	[52;59]	71	[70;73]
	$T_M o N_M$ (SR = 58 %)	1410	2156	567	2030	23	35	9	33	71	[69;73]	48	[47;50]	40	[38;41]	78	[77;80]
	$T_M u N_M$ (SR = 11 %)	397	311	1580	3875	6	5	26	63	20	[18;22]	93	[92;93]	56	[52;60]	71	[70;72]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-17 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 50 %)	T_M (SR = 57 %)	1,9	[1,81;2,00]	0,42	[0,39;0,45]	4,55	[4,08;5,07]	0,68	[0,66;0,69]	0,35	[0,32;0,39]	41
	N_M (SR = 13 %)	5,25	[4,36;6,31]	0,82	[0,80;0,84]	6,39	[5,24;7,80]	0,59	[0,58;0,60]	0,17	[0,15;0,19]	68
	$T_M o N_M$ (SR = 58 %)	1,91	[1,82;2,00]	0,4	[0,37;0,42]	4,82	[4,32;5,38]	0,68	[0,67;0,69]	0,36	[0,33;0,40]	43
	$T_M u N_M$ (SR = 11 %)	5,83	[4,77;7,13]	0,83	[0,82;0,85]	7,02	[5,67;8,69]	0,58	[0,57;0,60]	0,16	[0,14;0,18]	71
Verfehlen der Noten 1, 2 oder 3 (BR = 32 %)	T_M (SR = 57 %)	1,37	[1,31;1,43]	0,62	[0,58;0,67]	2,21	[1,98;2,48]	0,56	[0,55;0,57]	0,19	[0,15;0,22]	29
	N_M (SR = 13 %)	2,66	[2,34;3,04]	0,85	[0,83;0,87]	3,13	[2,68;3,65]	0,69	[0,68;0,70]	0,14	[0,11;0,16]	35
	$T_M o N_M$ (SR = 58 %)	1,38	[1,33;1,44]	0,59	[0,55;0,64]	2,34	[2,09;2,63]	0,56	[0,55;0,57]	0,2	[0,16;0,23]	32
	$T_M u N_M$ (SR = 11 %)	2,7	[2,35;3,10]	0,86	[0,84;0,88]	3,13	[2,67;3,67]	0,69	[0,68;0,70]	0,13	[0,10;0,15]	35

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-18 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 18 %)	T_D (SR = 30 %)	647	1226	468	3822	10	20	8	62	58	[55;61]	76	[75;77]	35	[32;37]	89	[88;90]
	N_D (SR = 7 %)	189	246	926	4802	3	4	15	78	17	[15;19]	95	[94;96]	43	[39;48]	84	[83;85]
	T_DoN_D (SR = 33 %)	684	1341	432	3706	11	22	7	60	61	[58;64]	73	[72;75]	34	[32;36]	90	[89;90]
	T_DuN_D (SR = 5 %)	152	131	963	4917	2	2	16	80	14	[12;16]	97	[97;98]	54	[48;60]	84	[83;85]
Verfehlen der Noten 1, 2 oder 3 (BR = 22 %)	T_D (SR = 30 %)	611	1262	718	3572	10	20	12	58	46	[43;49]	74	[73;75]	33	[31;35]	83	[82;84]
	N_D (SR = 7 %)	214	221	1114	4614	3	4	18	75	16	[14;18]	95	[95;96]	49	[44;54]	81	[80;82]
	T_DoN_D (SR = 33 %)	671	1354	658	3480	11	22	11	56	50	[48;53]	72	[71;73]	33	[31;35]	84	[83;85]
	T_DuN_D (SR = 5 %)	154	129	1174	4706	2	2	19	76	12	[10;13]	97	[97;98]	54	[48;60]	80	[79;81]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-18 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 18 %)	T_D (SR = 30 %)	2,39	[2,23;2,56]	0,55	[0,52;0,60]	4,31	[3,76;4,93]	0,73	[0,71;0,74]	0,34	[0,30;0,38]	40
	N_D (SR = 7 %)	3,48	[2,91;4,16]	0,87	[0,85;0,90]	3,98	[3,25;4,88]	0,81	[0,80;0,82]	0,12	[0,09;0,15]	31
	T_DoN_D (SR = 33 %)	2,31	[2,16;2,46]	0,53	[0,49;0,57]	4,38	[3,82;5,01]	0,71	[0,70;0,72]	0,35	[0,31;0,39]	42
	T_DuN_D (SR = 5 %)	5,25	[4,20;6,58]	0,89	[0,87;0,91]	5,92	[4,64;7,56]	0,82	[0,81;0,83]	0,11	[0,09;0,14]	43
Verfehlen der Noten 1, 2 oder 3 (BR = 22 %)	T_D (SR = 30 %)	1,76	[1,63;1,90]	0,73	[0,69;0,77]	2,41	[2,12;2,73]	0,68	[0,67;0,69]	0,2	[0,16;0,24]	22
	N_D (SR = 7 %)	3,53	[2,95;4,21]	0,88	[0,86;0,90]	4,01	[3,29;4,89]	0,78	[0,77;0,79]	0,12	[0,09;0,14]	35
	T_DoN_D (SR = 33 %)	1,8	[1,68;1,93]	0,69	[0,65;0,73]	2,62	[2,31;2,97]	0,67	[0,66;0,69]	0,22	[0,18;0,26]	26
	T_DuN_D (SR = 5 %)	4,35	[3,47;5,45]	0,91	[0,89;0,93]	4,79	[3,75;6,10]	0,79	[0,78;0,80]	0,09	[0,07;0,11]	42

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-19 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 5 %)	T_M (SR = 4 %)	46	129	204	4397	1	3	4	92	18	[14;24]	97	[97;98]	26	[20;33]	96	[95;96]
	N_M (SR = 0 %)	2	8	249	4517	0	0	5	95	1	[0;3]	100	[100;100]	20	[3;56]	95	[94;95]
	$T_M o N_M$ (SR = 4 %)	46	133	204	4393	1	3	4	92	18	[14;24]	97	[97;98]	26	[19;33]	96	[95;96]
	$T_M u N_M$ (SR = 0 %)	2	4	249	4521	0	0	5	95	1	[0;3]	100	[100;100]	33	[4;78]	95	[94;95]
Verfehlen der Noten 1, 2 oder 3 (BR = 26 %)	T_M (SR = 4 %)	86	89	1177	3424	2	2	25	72	7	[5;8]	97	[97;98]	49	[42;57]	74	[73;76]
	N_M (SR = 0 %)	4	6	1259	3507	0	0	26	73	0	[0;1]	100	[100;100]	40	[12;74]	74	[72;75]
	$T_M o N_M$ (SR = 4 %)	87	92	1176	3421	2	2	25	72	7	[6;8]	97	[97;98]	49	[41;56]	74	[73;76]
	$T_M u N_M$ (SR = 0 %)	3	3	1260	3510	0	0	26	73	0	[0;1]	100	[100;100]	50	[12;88]	74	[72;75]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-19 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 5 %)	T_M (SR = 4 %)	6,46	[4,73;8,82]	0,84	[0,79;0,89]	7,69	[5,34;11,07]	0,93	[0,92;0,94]	0,16	[0,10;0,21]	22
	N_M (SR = 0 %)	4,51	[0,96;21,11]	0,99	[0,98;1,00]	4,54	[0,96;21,47]	0,95	[0,94;0,95]	0,01	[0,00;0,03]	16
	$T_M o N_M$ (SR = 4 %)	6,26	[4,59;8,54]	0,84	[0,79;0,89]	7,45	[5,18;10,71]	0,93	[0,92;0,94]	0,15	[0,10;0,21]	22
	$T_M u N_M$ (SR = 0 %)	9,01	[1,66;48,98]	0,99	[0,98;1,00]	9,08	[1,65;49,80]	0,95	[0,94;0,95]	0,01	[0,00;0,03]	30
Verfehlen der Noten 1, 2 oder 3 (BR = 26 %)	T_M (SR = 4 %)	2,69	[2,01;3,59]	0,96	[0,94;0,97]	2,81	[2,07;3,81]	0,73	[0,72;0,75]	0,04	[0,02;0,06]	31
	N_M (SR = 0 %)	1,85	[0,52;6,56]	1	[1,00;1,00]	1,86	[0,52;6,59]	0,74	[0,72;0,75]	0	[0,00;0,01]	18
	$T_M o N_M$ (SR = 4 %)	2,63	[1,98;3,50]	0,96	[0,94;0,97]	2,75	[2,04;3,72]	0,73	[0,72;0,75]	0,04	[0,02;0,06]	30
	$T_M u N_M$ (SR = 0 %)	2,78	[0,56;13,76]	1	[1,00;1,00]	2,79	[0,56;13,82]	0,74	[0,72;0,75]	0	[0,00;0,01]	32

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-20 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 1 %)	T_D (SR = 1 %)	1	36	31	4708	0	1	1	99	3	[0;16]	99	[99;99]	3	[0;14]	99	[99;100]
	N_D (SR = 0 %)	1	5	31	4739	0	0	1	99	3	[0;16]	100	[100;100]	17	[0;64]	99	[99;100]
	T_DoN_D (SR = 1 %)	1	40	31	4704	0	1	1	98	3	[0;16]	99	[99;99]	2	[0;13]	99	[99;100]
	T_DuN_D (SR = 0 %)	1	1	31	4743	0	0	1	99	3	[0;16]	100	[100;100]	50	[1;99]	99	[99;100]
Verfehlen der Noten 1, 2 oder 3 (BR = 11 %)	T_D (SR = 1 %)	12	25	495	4244	0	1	10	89	2	[1;4]	99	[99;100]	32	[18;50]	90	[89;90]
	N_D (SR = 0 %)	3	3	504	4266	0	0	11	89	1	[0;2]	100	[100;100]	50	[12;88]	89	[89;90]
	T_DoN_D (SR = 1 %)	14	27	493	4242	0	1	10	89	3	[2;5]	99	[99;100]	34	[20;51]	90	[89;90]
	T_DuN_D (SR = 0 %)	1	1	505	4269	0	0	11	89	0	[0;1]	100	[100;100]	50	[1;99]	89	[89;90]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-20 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 1 %)	T_D (SR = 1 %)	4,12	[0,58;29,13]	0,98	[0,92;1,04]	4,22	[0,56;31,74]	0,99	[0,98;0,99]	0,02	[-0,01;0,16]	2
	N_D (SR = 0 %)	29,65	[3,56;246,70]	0,97	[0,91;1,03]	30,57	[3,47;269,36]	0,99	[0,99;0,99]	0,03	[0,00;0,16]	16
	$T_D o N_D$ (SR = 1 %)	3,71	[0,53;26,14]	0,98	[0,92;1,04]	3,79	[0,51;28,47]	0,99	[0,98;0,99]	0,02	[-0,01;0,16]	2
	$T_D u N_D$ (SR = 0 %)	148,25	[9,48;2318,84]	0,97	[0,91;1,03]	153	[9,36;2501,52]	0,99	[0,99;1,00]	0,03	[0,00;0,16]	50
Verfehlen der Noten 1, 2 oder 3 (BR = 11 %)	T_D (SR = 1 %)	4,04	[2,04;7,99]	0,98	[0,97;1,00]	4,12	[2,05;8,24]	0,89	[0,88;0,90]	0,02	[0,00;0,04]	24
	N_D (SR = 0 %)	8,42	[1,70;41,61]	0,99	[0,99;1,00]	8,46	[1,70;42,05]	0,89	[0,88;0,90]	0,01	[0,00;0,02]	44
	$T_D o N_D$ (SR = 1 %)	4,37	[2,30;8,27]	0,98	[0,96;0,99]	4,46	[2,32;8,57]	0,89	[0,88;0,90]	0,02	[0,01;0,04]	26
	$T_D u N_D$ (SR = 0 %)	8,44	[0,53;134,71]	1	[0,99;1,00]	8,45	[0,53;135,36]	0,89	[0,88;0,90]	0	[0,00;0,01]	44

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, $T_D o N_D$: Testergebnis in Deutsch oder Schulnote in Deutsch, $T_D u N_D$: Testergebnis in Deutsch und Schulnote in Deutsch.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-21 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 5 %)	T_M (SR = 16 %)	135	648	116	3877	3	14	2	81	54	[47;60]	86	[85;87]	17	[15;20]	97	[97;98]
	N_M (SR = 0 %)	2	8	249	4517	0	0	5	95	1	[0;3]	100	[100;100]	20	[3;56]	95	[94;95]
	$T_M o N_M$ (SR = 16 %)	135	650	116	3875	3	14	2	81	54	[47;60]	86	[85;87]	17	[15;20]	97	[97;98]
	$T_M u N_M$ (SR = 0 %)	2	6	249	4519	0	0	5	95	1	[0;3]	100	[100;100]	25	[3;65]	95	[94;95]
Verfehlen der Noten 1, 2 oder 3 (BR = 26 %)	T_M (SR = 16 %)	361	422	902	3091	8	9	19	65	29	[26;31]	88	[87;89]	46	[43;50]	77	[76;79]
	N_M (SR = 0 %)	4	6	1259	3507	0	0	26	73	0	[0;1]	100	[100;100]	40	[12;74]	74	[72;75]
	$T_M o N_M$ (SR = 16 %)	362	423	901	3090	8	9	19	65	29	[26;31]	88	[87;89]	46	[43;50]	77	[76;79]
	$T_M u N_M$ (SR = 0 %)	3	5	1260	3508	0	0	26	73	0	[0;1]	100	[100;100]	38	[9;76]	74	[72;75]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-21 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 5 %)	T_M (SR = 16 %)	3,76	[3,28;4,30]	0,54	[0,47;0,62]	6,96	[5,36;9,05]	0,84	[0,83;0,85]	0,39	[0,32;0,47]	45
	N_M (SR = 0 %)	4,51	[0,96;21,11]	0,99	[0,98;1,00]	4,54	[0,96;21,47]	0,95	[0,94;0,95]	0,01	[0,00;0,03]	16
	$T_M o N_M$ (SR = 16 %)	3,74	[3,27;4,29]	0,54	[0,47;0,62]	6,94	[5,34;9,01]	0,84	[0,83;0,85]	0,39	[0,32;0,47]	45
	$T_M u N_M$ (SR = 0 %)	6,01	[1,22;29,62]	0,99	[0,98;1,00]	6,05	[1,21;30,13]	0,95	[0,94;0,95]	0,01	[0,00;0,03]	21
Verfehlen der Noten 1, 2 oder 3 (BR = 26 %)	T_M (SR = 16 %)	2,38	[2,10;2,70]	0,81	[0,78;0,84]	2,93	[2,50;3,44]	0,72	[0,71;0,74]	0,17	[0,13;0,20]	27
	N_M (SR = 0 %)	1,85	[0,52;6,56]	1	[1,00;1,00]	1,86	[0,52;6,59]	0,74	[0,72;0,75]	0	[0,00;0,01]	18
	$T_M o N_M$ (SR = 16 %)	2,38	[2,10;2,70]	0,81	[0,78;0,84]	2,93	[2,50;3,44]	0,72	[0,71;0,74]	0,17	[0,13;0,20]	27
	$T_M u N_M$ (SR = 0 %)	1,67	[0,40;6,97]	1	[1,00;1,00]	1,67	[0,40;7,00]	0,74	[0,72;0,75]	0	[0,00;0,01]	15

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-22 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Mindeststandards (BR = 1 %)	T_D (SR = 4 %)	7	203	26	4540	0	4	1	95	21	[9;39]	96	[95;96]	3	[1;7]	99	[99;100]
	N_D (SR = 0 %)	1	5	31	4739	0	0	1	99	3	[0;16]	100	[100;100]	17	[0;64]	99	[99;100]
	T_DoN_D (SR = 4 %)	7	207	25	4537	0	4	1	95	22	[9;40]	96	[95;96]	3	[1;7]	99	[99;100]
	T_DuN_D (SR = 0 %)	1	1	31	4743	0	0	1	99	3	[0;16]	100	[100;100]	50	[1;99]	99	[99;100]
Verfehlen der Noten 1, 2 oder 3 (BR = 11 %)	T_D (SR = 4 %)	47	163	459	4107	1	3	10	86	9	[7;12]	96	[96;97]	22	[17;29]	90	[89;91]
	N_D (SR = 0 %)	3	3	504	4266	0	0	11	89	1	[0;2]	100	[100;100]	50	[12;88]	89	[89;90]
	T_DoN_D (SR = 4 %)	49	165	458	4104	1	3	10	86	10	[7;13]	96	[96;97]	23	[17;29]	90	[89;91]
	T_DuN_D (SR = 0 %)	1	1	505	4269	0	0	11	89	0	[0;1]	100	[100;100]	50	[1;99]	89	[89;90]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-22 (Teil B)

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

Kriterium 8. Klasse	Prädiktor(en) 3. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Mindeststandards (BR = 1 %)	T_D (SR = 4 %)	4,96	[2,53;9,70]	0,82	[0,69;0,98]	6,02	[2,58;14,04]	0,95	[0,95;0,96]	0,17	[0,04;0,35]	18
	N_D (SR = 0 %)	29,65	[3,56;246,70]	0,97	[0,91;1,03]	30,57	[3,47;269,36]	0,99	[0,99;0,99]	0,03	[0,00;0,16]	16
	T_DoN_D (SR = 4 %)	5,01	[2,57;9,78]	0,82	[0,68;0,98]	6,14	[2,62;14,35]	0,95	[0,94;0,96]	0,18	[0,04;0,36]	18
	T_DuN_D (SR = 0 %)	148,25	[9,48;2318,84]	0,97	[0,91;1,03]	153	[9,36;2501,52]	0,99	[0,99;1,00]	0,03	[0,00;0,16]	50
Verfehlen der Noten 1, 2 oder 3 (BR = 11 %)	T_D (SR = 4 %)	2,43	[1,78;3,32]	0,94	[0,92;0,97]	2,58	[1,84;3,62]	0,87	[0,86;0,88]	0,05	[0,02;0,09]	13
	N_D (SR = 0 %)	8,42	[1,70;41,61]	0,99	[0,99;1,00]	8,46	[1,70;42,05]	0,89	[0,88;0,90]	0,01	[0,00;0,02]	44
	T_DoN_D (SR = 4 %)	2,5	[1,84;3,39]	0,94	[0,91;0,97]	2,66	[1,91;3,71]	0,87	[0,86;0,88]	0,06	[0,03;0,09]	14
	T_DuN_D (SR = 0 %)	8,44	[0,53;134,71]	1	[0,99;1,00]	8,45	[0,53;135,36]	0,89	[0,88;0,90]	0	[0,00;0,01]	44

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_D: Testergebnis in Deutsch, N_D: Schulnote in Deutsch, T_DoN_D: Testergebnis in Deutsch oder Schulnote in Deutsch, T_DuN_D: Testergebnis in Deutsch und Schulnote in Deutsch.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-23 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 4. Klasse. Prognose von der 4. Klasse zur 5. Klasse.

Kriterium 5. Klasse	Prädiktor(en) 4. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Regelstandards (der 4. Klasse) (BR = 17 %)	T_M (SR = 8 %)	24	10	48	348	6	2	11	81	33	[23;45]	97	[95;99]	71	[53;85]	88	[84;91]
	N_M (SR = 8 %)	24	12	48	346	6	3	11	80	33	[23;45]	97	[94;98]	67	[49;81]	88	[84;91]
	$T_M o N_M$ (SR = 13 %)	36	20	36	338	8	5	8	79	50	[38;62]	94	[92;97]	64	[50;77]	90	[87;93]
	$T_M u N_M$ (SR = 3 %)	12	2	60	356	3	0	14	83	17	[9;27]	99	[98;100]	86	[57;98]	86	[82;89]
Verfehlen der Noten 1, 2 oder 3 (BR = 15 %)	T_M (SR = 8 %)	18	16	46	350	4	4	11	81	28	[18;41]	96	[93;97]	53	[35;70]	88	[85;91]
	N_M (SR = 8 %)	27	9	36	358	6	2	8	83	43	[30;56]	98	[95;99]	75	[58;88]	91	[88;94]
	$T_M o N_M$ (SR = 13 %)	33	23	31	343	8	5	7	80	52	[39;64]	94	[91;96]	59	[45;72]	92	[88;94]
	$T_M u N_M$ (SR = 3 %)	13	1	51	365	3	0	12	85	20	[11;32]	100	[98;100]	93	[66;100]	88	[84;91]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-23 (Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 4. Klasse. Prognose von der 4. Klasse zur 5. Klasse.

Kriterium 5. Klasse	Prädiktor(en) 4. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Regelstandards (der 4. Klasse) (BR = 17 %)	T_M (SR = 8 %)	12	[6;24]	0,69	[0,58;0,81]	17,4	[7,84;38,61]	0,87	[0,83;0,90]	0,31	[0,18;0,44]	65
	N_M (SR = 8 %)	10	[5;19]	0,69	[0,59;0,81]	14,42	[6,77;30,70]	0,86	[0,82;0,89]	0,3	[0,17;0,44]	60
	$T_M o N_M$ (SR = 13 %)	9	[6;15]	0,53	[0,42;0,67]	16,9	[8,86;32,23]	0,87	[0,83;0,90]	0,44	[0,29;0,59]	57
	$T_M u N_M$ (SR = 3 %)	30	[7;130]	0,84	[0,76;0,93]	35,6	[7,77;163,05]	0,86	[0,82;0,89]	0,16	[0,07;0,27]	83
Verfehlen der Noten 1, 2 oder 3 (BR = 15 %)	T_M (SR = 8 %)	6	[3;12]	0,75	[0,64;0,88]	8,56	[4,08;17,95]	0,86	[0,82;0,89]	0,24	[0,11;0,38]	45
	N_M (SR = 8 %)	17	[9;35]	0,59	[0,47;0,73]	29,83	[13,03;68,32]	0,9	[0,86;0,92]	0,4	[0,26;0,55]	71
	$T_M o N_M$ (SR = 13 %)	8	[5;13]	0,52	[0,40;0,67]	15,88	[8,31;30,32]	0,87	[0,84;0,90]	0,45	[0,29;0,60]	52
	$T_M u N_M$ (SR = 3 %)	74	[10;558]	0,8	[0,71;0,90]	93,04	[11,92;726,30]	0,88	[0,84;0,91]	0,2	[0,10;0,32]	92

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-24 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 4. Klasse. Prognose von der 4. Klasse zur 6. Klasse.

Kriterium 6. Klasse	Prädiktor(en) 4. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Regelstandards (der 4. Klasse) (BR = 14 %)	T_M (SR = 8 %)	17	17	45	351	4	4	10	82	27	[17;40]	95	[93;97]	50	[32;68]	89	[85;92]
	N_M (SR = 8 %)	22	14	40	354	5	3	9	82	35	[24;49]	96	[94;98]	61	[43;77]	90	[86;93]
	$T_M o N_M$ (SR = 13 %)	30	26	32	342	7	6	7	80	48	[35;61]	93	[90;95]	54	[40;67]	91	[88;94]
	$T_M u N_M$ (SR = 3 %)	8	6	54	362	2	1	13	84	13	[6;24]	98	[96;99]	57	[29;82]	87	[83;90]
Verfehlen der Noten 1, 2 oder 3 (BR = 18 %)	T_M (SR = 8 %)	21	13	56	340	5	3	13	79	27	[18;39]	96	[94;98]	62	[44;78]	86	[82;89]
	N_M (SR = 8 %)	28	8	49	345	7	2	11	80	36	[26;48]	98	[96;99]	78	[61;90]	88	[84;91]
	$T_M o N_M$ (SR = 13 %)	35	21	42	332	8	5	10	77	45	[34;57]	94	[91;96]	63	[49;75]	89	[85;92]
	$T_M u N_M$ (SR = 3 %)	14	0	63	353	3	0	15	82	18	[10;29]	100	[99;100]	100	[77;100]	85	[81;88]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-24 (Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 4. Klasse. Prognose von der 4. Klasse zur 6. Klasse.

Kriterium 6. Klasse	Prädiktor(en) 4. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Regelstandards (der 4. Klasse) (BR = 14 %)	T_M (SR = 8 %)	5,94	[3,21;10,99]	0,76	[0,65;0,89]	7,8	[3,72;16,36]	0,86	[0,82;0,89]	0,23	[0,10;0,38]	42
	N_M (SR = 8 %)	9,33	[5,05;17,23]	0,67	[0,56;0,81]	13,91	[6,60;29,31]	0,87	[0,84;0,90]	0,32	[0,17;0,47]	55
	$T_M o N_M$ (SR = 13 %)	6,85	[4,36;10,75]	0,56	[0,44;0,71]	12,33	[6,52;23,34]	0,87	[0,83;0,90]	0,41	[0,25;0,57]	46
	$T_M u N_M$ (SR = 3 %)	7,91	[2,84;22,03]	0,89	[0,80;0,98]	8,94	[2,99;26,76]	0,86	[0,82;0,89]	0,11	[0,02;0,23]	50
Verfehlen der Noten 1, 2 oder 3 (BR = 18 %)	T_M (SR = 8 %)	7,41	[3,88;14,13]	0,76	[0,66;0,87]	9,81	[4,65;20,70]	0,84	[0,80;0,87]	0,24	[0,12;0,37]	53
	N_M (SR = 8 %)	16,05	[7,61;33,83]	0,65	[0,55;0,77]	24,64	[10,63;57,13]	0,87	[0,83;0,90]	0,34	[0,21;0,47]	73
	$T_M o N_M$ (SR = 13 %)	7,64	[4,72;12,37]	0,58	[0,47;0,71]	13,17	[7,02;24,71]	0,85	[0,82;0,89]	0,4	[0,25;0,53]	54
	$T_M u N_M$ (SR = 3 %)	-	[-;-]	0,82	[0,74;0,91]	-	[-;-]	0,85	[0,82;0,89]	0,18	[0,09;0,29]	100

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-25 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 4. Klasse. Prognose von der 4. Klasse zur 5. Klasse.

Kriterium 5. Klasse	Prädiktor(en) 4. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Regelstandards (der 4. Klasse) (BR = 17 %)	T_M (SR = 27 %)	50	65	22	293	12	15	5	68	69	[57;80]	82	[77;86]	43	[34;53]	93	[90;96]
	N_M (SR = 8 %)	24	12	48	346	6	3	11	80	33	[23;45]	97	[94;98]	67	[49;81]	88	[84;91]
	$T_M o N_M$ (SR = 29 %)	56	69	16	289	13	16	4	67	78	[66;87]	81	[76;85]	45	[36;54]	95	[92;97]
	$T_M u N_M$ (SR = 6 %)	18	8	54	350	4	2	13	81	25	[16;37]	98	[96;99]	69	[48;86]	87	[83;90]
Verfehlen der Noten 1, 2 oder 3 (BR = 15 %)	T_M (SR = 27 %)	41	74	23	292	10	17	5	68	64	[51;76]	80	[75;84]	36	[27;45]	93	[89;95]
	N_M (SR = 8 %)	27	9	36	358	6	2	8	83	43	[30;56]	98	[95;99]	75	[58;88]	91	[88;94]
	$T_M o N_M$ (SR = 29 %)	48	77	16	289	11	18	4	67	75	[63;85]	79	[74;83]	38	[30;48]	95	[92;97]
	$T_M u N_M$ (SR = 6 %)	21	5	43	361	5	1	10	84	33	[22;46]	99	[97;100]	81	[61;93]	89	[86;92]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-25 (Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 4. Klasse. Prognose von der 4. Klasse zur 5. Klasse.

Kriterium 5. Klasse	Prädiktor(en) 4. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Regelstandards (der 4. Klasse) (BR = 17 %)	T_M (SR = 27 %)	4	[3;5]	0,37	[0,26;0,53]	10,24	[5,80;18,10]	0,8	[0,76;0,83]	0,51	[0,35;0,65]	58
	N_M (SR = 8 %)	10	[5;19]	0,69	[0,59;0,81]	14,42	[6,77;30,70]	0,86	[0,82;0,89]	0,3	[0,17;0,44]	60
	$T_M o N_M$ (SR = 29 %)	4	[3;5]	0,28	[0,18;0,43]	14,66	[7,93;27,10]	0,8	[0,76;0,84]	0,59	[0,43;0,71]	69
	$T_M u N_M$ (SR = 6 %)	11	[5;25]	0,77	[0,67;0,88]	14,58	[6,04;35,19]	0,86	[0,82;0,89]	0,23	[0,11;0,36]	63
Verfehlen der Noten 1, 2 oder 3 (BR = 15 %)	T_M (SR = 27 %)	3	[2;4]	0,45	[0,32;0,63]	7,03	[3,97;12,45]	0,77	[0,73;0,81]	0,44	[0,26;0,59]	51
	N_M (SR = 8 %)	17	[9;35]	0,59	[0,47;0,73]	29,83	[13,03;68,32]	0,9	[0,86;0,92]	0,4	[0,26;0,55]	71
	$T_M o N_M$ (SR = 29 %)	4	[3;5]	0,32	[0,21;0,49]	11,26	[6,06;20,91]	0,78	[0,74;0,82]	0,54	[0,37;0,68]	65
	$T_M u N_M$ (SR = 6 %)	24	[9;61]	0,68	[0,57;0,81]	35,26	[12,65;98,30]	0,89	[0,85;0,92]	0,31	[0,18;0,45]	77

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

Tabelle ESM-26 (Teil A – Fortsetzung der Tabelle auf nachfolgender Seite mit Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 4. Klasse. Prognose von der 4. Klasse zur 6. Klasse.

Kriterium 6. Klasse	Prädiktor(en) 4. Klasse	RP	FP	FN	RN	RP (%)	FP (%)	FN (%)	RN (%)	Sen (in %)		Spe (in %)		ppW (in %)		npW (in %)	
										Par	KI	Par	KI	Par	KI	Par	KI
Verfehlen des Regelstandards (der 4. Klasse) (BR = 14 %)	T_M (SR = 27 %)	38	77	24	291	9	18	6	68	61	[48;73]	79	[75;83]	33	[25;42]	92	[89;95]
	N_M (SR = 8 %)	22	14	40	354	5	3	9	82	35	[24;49]	96	[94;98]	61	[43;77]	90	[86;93]
	T_MoN_M (SR = 29 %)	44	81	18	287	10	19	4	67	71	[58;82]	78	[73;82]	35	[27;44]	94	[91;96]
	T_MuN_M (SR = 6 %)	16	10	46	358	4	2	11	83	26	[16;38]	97	[95;99]	62	[41;80]	89	[85;92]
Verfehlen der Noten 1, 2 oder 3 (BR = 18 %)	T_M (SR = 27 %)	44	71	33	282	10	17	8	66	57	[45;68]	80	[75;84]	38	[29;48]	90	[86;93]
	N_M (SR = 8 %)	28	8	49	345	7	2	11	80	36	[26;48]	98	[96;99]	78	[61;90]	88	[84;91]
	T_MoN_M (SR = 29 %)	51	74	26	279	12	17	6	65	66	[55;77]	79	[74;83]	41	[32;50]	91	[88;94]
	T_MuN_M (SR = 6 %)	21	5	56	348	5	1	13	81	27	[18;39]	99	[97;100]	81	[61;93]	86	[82;89]

Anmerkungen. RP: Richtig Positive, FP: Falsch Positive, FN: Falsch Negative, RN: Richtig Negative, Sen: Sensitivität, Spe: Spezifität, ppW: positiv prädiktiver Wert, npW: negativ prädiktiver Wert, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, T_MoN_M: Testergebnis in Mathematik oder Schulnote in Mathematik, T_MuN_M: Testergebnis in Mathematik und Schulnote in Mathematik

Tabelle ESM-26 (Teil B)

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall – Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 4. Klasse. Prognose von der 4. Klasse zur 6. Klasse.

Kriterium 6. Klasse	Prädiktor(en) 4. Klasse	+LR		-LR		Odds Ratio		GTQ		Youden-Index		RATZ[a] (in %)
		Par	KI	Par	KI	Par	KI	Par	KI	Par	KI	Par
Verfehlen des Regelstandards (der 4. Klasse) (BR = 14 %)	T_M (SR = 27 %)	2,93	[2,21;3,88]	0,49	[0,36;0,67]	5,98	[3,39;10,57]	0,77	[0,72;0,80]	0,4	[0,23;0,57]	47
	N_M (SR = 8 %)	9,33	[5,05;17,23]	0,67	[0,56;0,81]	13,91	[6,60;29,31]	0,87	[0,84;0,90]	0,32	[0,17;0,47]	55
	$T_M o N_M$ (SR = 29 %)	3,22	[2,51;4,14]	0,37	[0,25;0,55]	8,66	[4,75;15,80]	0,77	[0,73;0,81]	0,49	[0,31;0,64]	59
	$T_M u N_M$ (SR = 6 %)	9,5	[4,52;19,96]	0,76	[0,66;0,88]	12,45	[5,33;29,06]	0,87	[0,83;0,90]	0,23	[0,11;0,37]	55
Verfehlen der Noten 1, 2 oder 3 (BR = 18 %)	T_M (SR = 27 %)	2,84	[2,14;3,77]	0,54	[0,41;0,70]	5,3	[3,15;8,92]	0,76	[0,71;0,80]	0,37	[0,21;0,52]	41
	N_M (SR = 8 %)	16,05	[7,61;33,83]	0,65	[0,55;0,77]	24,64	[10,63;57,13]	0,87	[0,83;0,90]	0,34	[0,21;0,47]	73
	$T_M o N_M$ (SR = 29 %)	3,16	[2,44;4,09]	0,43	[0,31;0,59]	7,4	[4,32;12,66]	0,77	[0,72;0,81]	0,45	[0,29;0,60]	52
	$T_M u N_M$ (SR = 6 %)	19,25	[7,49;49,47]	0,74	[0,64;0,85]	26,1	[9,46;72,04]	0,86	[0,82;0,89]	0,26	[0,14;0,38]	77

Anmerkungen. +LR: Positiver Likelihood Ratio, -LR: Negativer Likelihood Ratio, GTQ: Gesamttrefferquote, Par: Parameter, KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

[a] Berechnung nach Tröster (2009, S. 143).

ESM-15 bis 26: Literatur

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

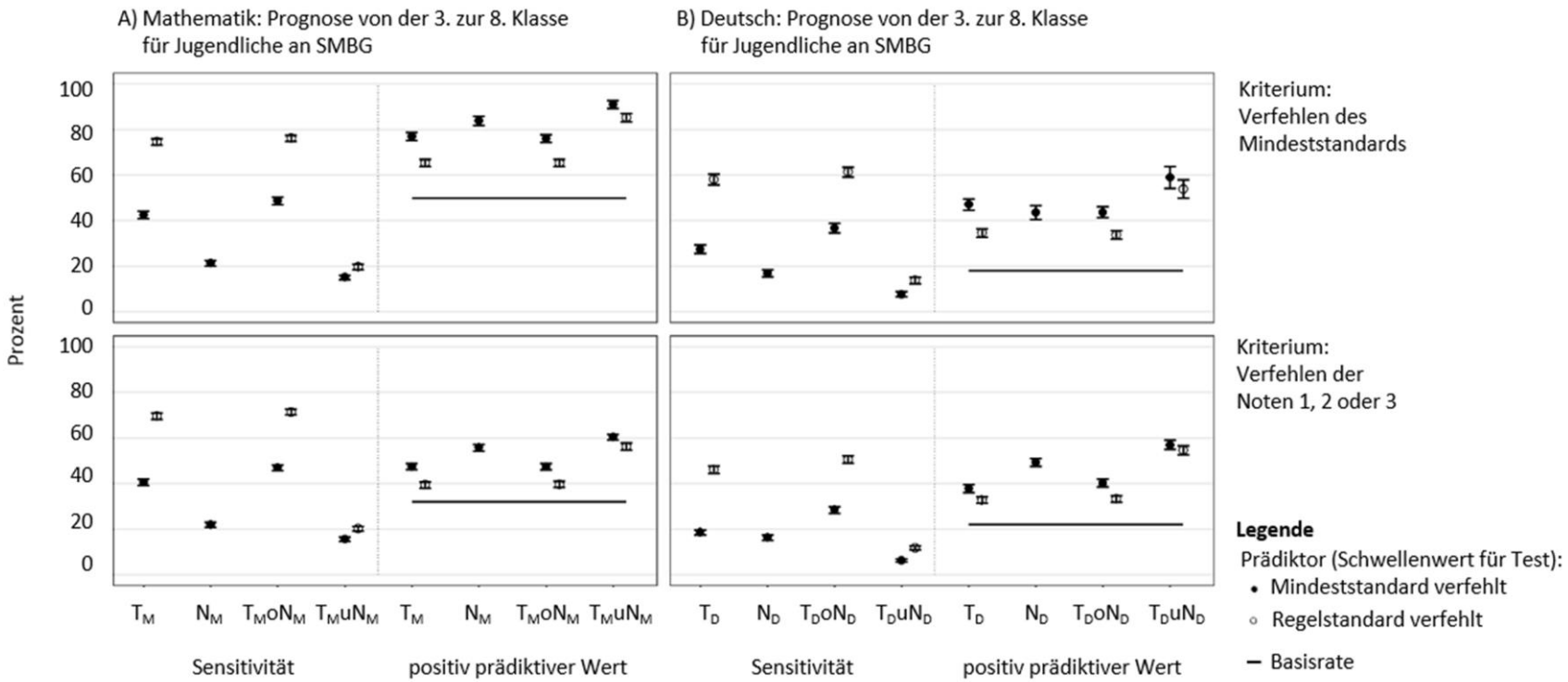


Abbildung ESM-27. Ergebnisse mit 95%-Konfidenzintervallen nach Rubin (1987)[a]. Hauptstudie - Prognosen für Jugendliche an SMBG in (A) Mathematik und (B) Deutsch: Sensitivität (*Sen*) und positiv prädiktive Werte (*ppW*, in Prozent) für das Verfehlen des Mindeststandards (= unter Mindeststandard) und das Verfehlen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Strategien (Spezifikation s. Text) und (2) der Schwelle für die Tests (Verfehlen der Mindest- bzw. Regelstandards). T_M bzw. T_D = Testergebnis in Mathematik bzw. Deutsch, N_M bzw. N_D = Schulnote in Mathematik bzw. Deutsch, T_MoN_M bzw. T_DoN_D = Testergebnis oder Note in Mathematik bzw. Deutsch, T_MuN_M bzw. T_DuN_D = Testergebnis und Note in Mathematik bzw. Deutsch. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Jugendlichen, die in der 8. Klasse ein bestimmtes mathematik- bzw. deutschbezogenes Bildungsergebnis verfehlten).

[a] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

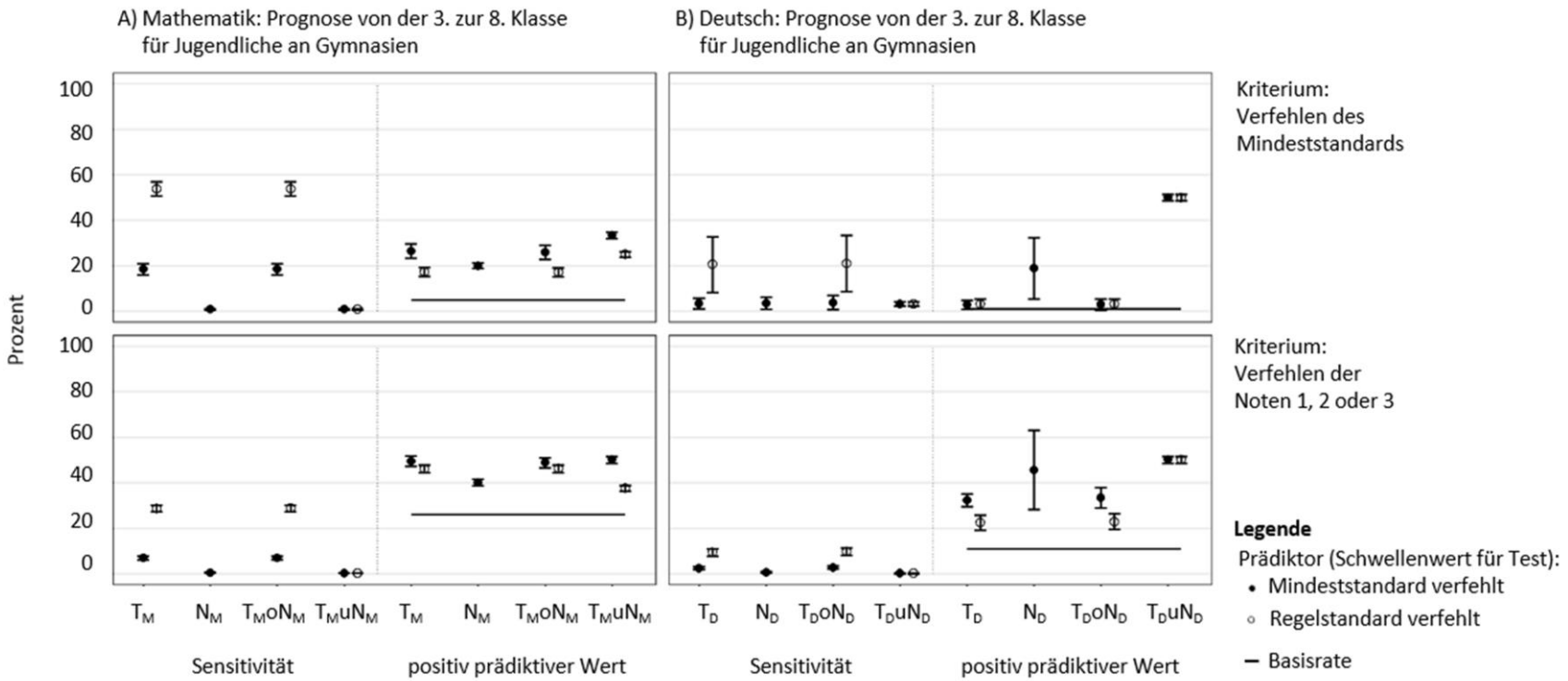


Abbildung ESM-28. Ergebnisse mit 95%-Konfidenzintervallen nach Rubin (1987)[a]. Hauptstudie - Prognosen für Jugendliche an Gymnasien in (A) Mathematik und (B) Deutsch: Sensitivität (*Sen*) und positiv prädiktive Werte (*ppW*, in Prozent) für das Verfehlen des Mindeststandards (= unter Mindeststandard) und das Verfehlen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Strategien (Spezifikation s. Text) und (2) der Schwelle für die Tests (Verfehlen der Mindest- bzw. Regelstandards). T_M bzw. T_D = Testergebnis in Mathematik bzw. Deutsch, N_M bzw. N_D = Schulnote in Mathematik bzw. Deutsch, T_MoN_M bzw. T_DoN_D = Testergebnis oder Note in Mathematik bzw. Deutsch, T_MuN_M bzw. T_DuN_D = Testergebnis und Note in Mathematik bzw. Deutsch. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Jugendlichen, die in der 8. Klasse ein bestimmtes mathematik- bzw. deutschbezogenes Bildungsergebnis verfehlten).

[a] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

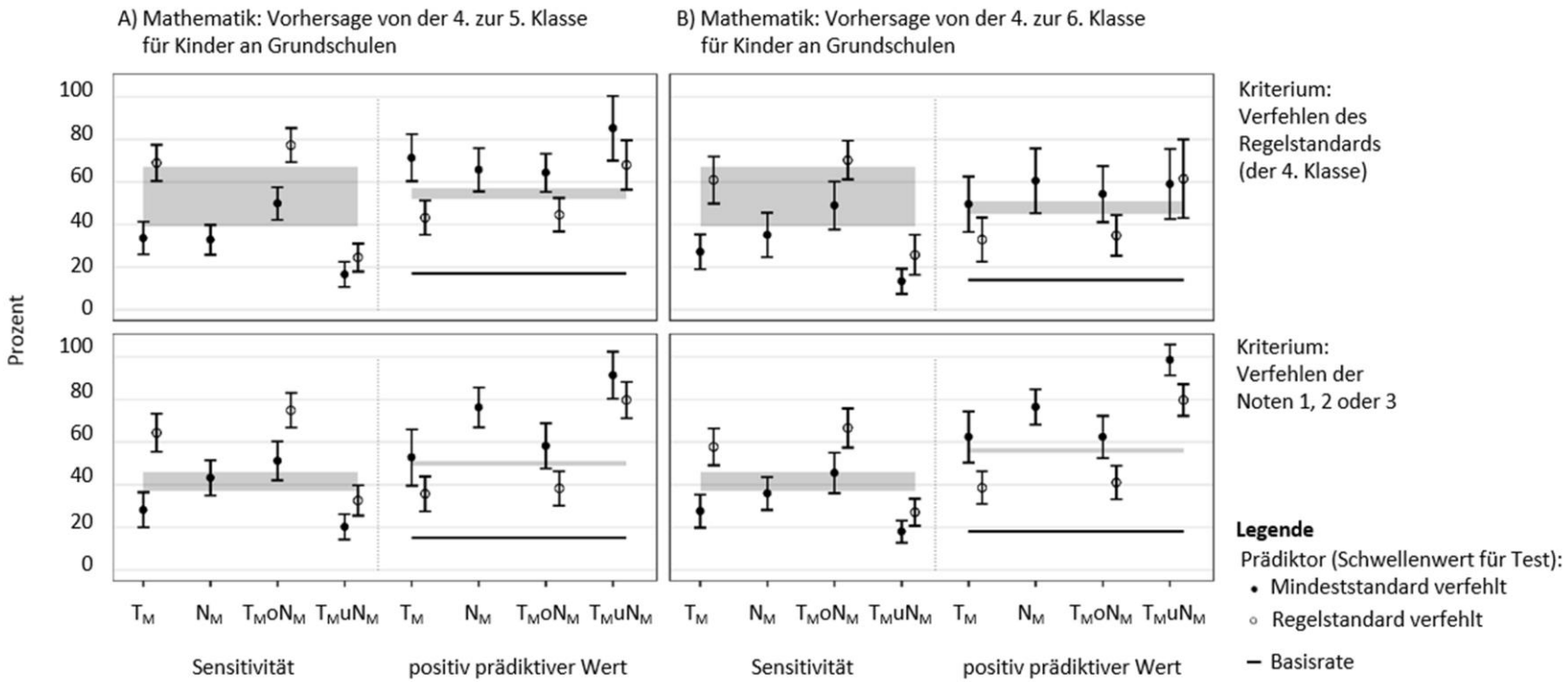


Abbildung ESM-29. Ergebnisse mit 95%-Konfidenzintervallen nach Rubin (1987)[a]. Ergänzungsstudie - Ergebnisse für Kinder an Grundschulen im Fach Mathematik für (A) Prognosen von der 4. zur 5. Klasse und (B) Prognosen von der 4. zur 6. Klasse: Sensitivität (*Sen*) und positiv prädiktive Werte (*ppW*, in Prozent) für das Verfehlen des Regelstandards (= unter Regelstandard) und das Verfehlen der Schulnoten 1, 2 oder 3 (= Note 4, 5 oder 6) in Abhängigkeit von (1) den verwendeten Prädiktoren bzw. Strategien (Spezifikation s. Text) und (2) der Schwelle für die Tests (Verfehlen der Mindest- bzw. Regelstandards). T_M = Testergebnis in Mathematik, N_M = Schulnote in Mathematik, T_MoN_M = Testergebnis oder Note in Mathematik, T_MuN_M = Testergebnis und Note in Mathematik. Die horizontalen Linien bei den positiv prädiktiven Werten markieren die jeweilige Basisrate (= Anteil der Kinder, die in der 5. bzw. 6. Klasse ein bestimmtes mathematikbezogenes Bildungsergebnis verfehlten). Die grau unterlegten Werte markieren den Referenzbereich, der für deutschsprachige standardisierte Schulleistungstests ermittelt wurde (für Details s. ESM-36).

[a] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

ESM-27 bis 29: Literatur

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Tabelle ESM-30

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse Mathematik/Deutsch	Prädiktor(en) 3. Klasse Mathematik/Deutsch	Sensitivität (in %)		positiv prädiktiver Wert (in %)		Sensitivität (in %)		positiv prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Verfehlen des Mindeststandards (BR = 50 %/18 %)	T_M/T_D (SR = 27 %/11 %)	42	[41;44]	77	[75;79]	27	[25;30]	47	[43;51]
	N_M/N_D (SR = 13 %/7 %)	21	[20;23]	84	[81;86]	17	[15;19]	43	[39;48]
	$T_M o N_M/T_D o N_D$ (SR = 32 %/15 %)	49	[47;50]	76	[74;78]	37	[34;40]	44	[40;47]
	$T_M u N_M/T_D u N_D$ (SR = 8 %/2 %)	15	[14;16]	91	[88;93]	8	[6;9]	59	[50;67]
Verfehlen der Noten 1, 2 oder 3 (BR = 32 %/22 %)	T_M/T_D (SR = 27 %/11 %)	40	[38;43]	47	[45;50]	18	[16;21]	38	[34;42]
	N_M/N_D (SR = 13 %/7 %)	22	[20;24]	56	[52;59]	16	[14;18]	49	[44;54]
	$T_M o N_M/T_D o N_D$ (SR = 32 %/15 %)	47	[45;49]	47	[45;50]	28	[26;31]	40	[37;43]
	$T_M u N_M/T_D u N_D$ (SR = 8 %/2 %)	15	[14;17]	60	[56;65]	6	[5;8]	57	[48;65]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, $T_M/_{TD}$: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, $T_M o N_M/T_D o N_D$: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, $T_M u N_M/T_D u N_D$: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-31

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Schulen mit mehreren Bildungsgängen (SMBG) im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse Mathematik/Deutsch	Prädiktor(en) 3. Klasse Mathematik/Deutsch	Sensitivität (in %)		positiv prädiktiver Wert (in %)		Sensitivität (in %)		positiv prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Verfehlen des Mindeststandards (BR = 50 %/18 %)	T_M/T_D (SR = 57 %/30 %)	75	[73;76]	65	[64;67]	58	[55;61]	35	[32;37]
	N_M/N_D (SR = 13 %/7 %)	21	[20;23]	84	[81;86]	17	[15;19]	43	[39;48]
	$T_M o N_M/T_D o N_D$ (SR = 58 %/33 %)	76	[75;78]	65	[64;67]	61	[58;64]	34	[32;36]
	$T_M u N_M/T_D u N_D$ (SR = 11 %/5 %)	20	[18;21]	85	[82;88]	14	[12;16]	54	[48;60]
Verfehlen der Noten 1, 2 oder 3 (BR = 32 %/22 %)	T_M/T_D (SR = 57 %/30 %)	69	[67;72]	39	[38;41]	46	[43;49]	33	[31;35]
	N_M/N_D (SR = 13 %/7 %)	22	[20;24]	56	[52;59]	16	[14;18]	49	[44;54]
	$T_M o N_M/T_D o N_D$ (SR = 58 %/33 %)	71	[69;73]	40	[38;41]	50	[48;53]	33	[31;35]
	$T_M u N_M/T_D u N_D$ (SR = 11 %/5 %)	20	[18;22]	56	[52;60]	12	[10;13]	54	[48;60]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M/T_D: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, $T_M o N_M/T_D o N_D$: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, $T_M u N_M/T_D u N_D$: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-32

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse Mathematik/Deutsch	Prädiktor(en) 3. Klasse Mathematik/Deutsch	Sensitivität (in %)		positiv prädiktiver Wert (in %)		Sensitivität (in %)		positiv prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Verfehlen des Mindeststandards (BR = 5 %/1 %)	T_M/T_D (SR = 4 %/1 %)	18	[14;24]	26	[20;33]	3	[0;16]	3	[0;14]
	N_M/N_D (SR = 0 %/0 %)	1	[0;3]	20	[3;56]	3	[0;16]	17	[0;64]
	T_MoN_M/T_DoN_D (SR = 4 %/1 %)	18	[14;24]	26	[19;33]	3	[0;16]	2	[0;13]
	T_MuN_M/T_DuN_D (SR = 0 %/0 %)	1	[0;3]	33	[4;78]	3	[0;16]	50	[1;99]
Verfehlen der Noten 1, 2 oder 3 (BR = 26 %/11 %)	T_M/T_D (SR = 4 %/1 %)	7	[5;8]	49	[42;57]	2	[1;4]	32	[18;50]
	N_M/N_D (SR = 0 %/0 %)	0	[0;1]	40	[12;74]	1	[0;2]	50	[12;88]
	T_MoN_M/T_DoN_D (SR = 4 %/1 %)	7	[6;8]	49	[41;56]	3	[2;5]	34	[20;51]
	T_MuN_M/T_DuN_D (SR = 0 %/0 %)	0	[0;1]	50	[12;88]	0	[0;1]	50	[1;99]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M/T_D: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, T_MoN_M/T_DoN_D: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, T_MuN_M/T_DuN_D: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-33

Ergebnisse der Hauptstudie für Prognosen von der 3. zur 8. Klasse für Jugendliche an Gymnasien im Fach (A) Mathematik und (B) Deutsch: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 3. Klasse

		A) Mathematik				B) Deutsch			
Kriterium 8. Klasse Mathematik/Deutsch	Prädiktor(en) 3. Klasse Mathematik/Deutsch	Sensitivität (in %)		positiv prädiktiver Wert (in %)		Sensitivität (in %)		positiv prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Verfehlen des Mindeststandards (BR = 5 %/1 %)	T_M/T_D (SR = 16 %/4 %)	54	[47;60]	17	[15;20]	21	[9;39]	3	[1;7]
	N_M/N_D (SR = 0 %/0 %)	1	[0;3]	20	[3;56]	3	[0;16]	17	[0;64]
	T_MoN_M/T_DoN_D (SR = 16 %/4 %)	54	[47;60]	17	[15;20]	22	[9;40]	3	[1;7]
	T_MuN_M/T_DuN_D (SR = 0 %/0 %)	1	[0;3]	25	[3;65]	3	[0;16]	50	[1;99]
Verfehlen der Noten 1, 2 oder 3 (BR = 26 %/11 %)	T_M/T_D (SR = 16 %/4 %)	29	[26;31]	46	[43;50]	9	[7;12]	22	[17;29]
	N_M/N_D (SR = 0 %/0 %)	0	[0;1]	40	[12;74]	1	[0;2]	50	[12;88]
	T_MoN_M/T_DoN_D (SR = 16 %/4 %)	29	[26;31]	46	[43;50]	10	[7;13]	23	[17;29]
	T_MuN_M/T_DuN_D (SR = 0 %/0 %)	0	[0;1]	38	[9;76]	0	[0;1]	50	[1;99]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M/T_D: Testergebnis in Mathematik/Testergebnis in Deutsch, N_M/N_D: Schulnote in Mathematik/Schulnote in Deutsch, T_MoN_M/T_DoN_D: Testergebnis in Mathematik oder Schulnote in Mathematik/Testergebnis in Deutsch oder Schulnote in Deutsch, T_MuN_M/T_DuN_D: Testergebnis in Mathematik und Schulnote in Mathematik/Testergebnis in Deutsch und Schulnote in Deutsch.

Tabelle ESM-34

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Mindeststandards in 4. Klasse

		Prognose von der 4. Klasse zur 5. Klasse				Prognose von der 4. Klasse zur 6. Klasse			
Kriterium 5. Klasse/6. Klasse	Prädiktor(en) 4. Klasse	Sensitivität (in %)		positiv prädiktiver Wert (in %)		Sensitivität (in %)		positiv prädiktiver Wert (in %)	
Mathematik	Mathematik	Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Verfehlen des Regelstandards (der 4. Klasse) (BR = 17 %/14 %)	T_M (SR = 8 %/8 %)	33	[23;45]	71	[53;85]	27	[17;40]	50	[32;68]
	N_M (SR = 8 %/8 %)	33	[23;45]	67	[49;81]	35	[24;49]	61	[43;77]
	$T_M o N_M$ (SR = 13 %/13 %)	50	[38;62]	64	[50;77]	48	[35;61]	54	[40;67]
	$T_M u N_M$ (SR = 3 %/3 %)	17	[9;27]	86	[57;98]	13	[6;24]	57	[29;82]
Verfehlen der Noten 1, 2 oder 3(BR = 15 %/18 %)	T_M (SR = 8 %/8 %)	28	[18;41]	53	[35;70]	27	[18;39]	62	[44;78]
	N_M (SR = 8 %/8 %)	43	[30;56]	75	[58;88]	36	[26;48]	78	[61;90]
	$T_M o N_M$ (SR = 13 %/13 %)	52	[39;64]	59	[45;72]	45	[34;57]	63	[49;75]
	$T_M u N_M$ (SR = 3 %/3 %)	20	[11;32]	93	[66;100]	18	[10;29]	100	[77;100]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

Tabelle ESM-35

Ergebnisse für die Ergänzungsstudie im Fach Mathematik: Klassifikatorische Indizes zur Prognosegüte mit 95%-Konfidenzintervall — Identifikation „gefährdeter“ Kinder durch Verfehlen des Regelstandards in 4. Klasse

		Prognose von der 4. Klasse zur 5. Klasse				Prognose von der 4. Klasse zur 6. Klasse			
Kriterium 5. Klasse/6. Klasse Mathematik	Prädiktor(en) 4. Klasse Mathematik	Sensitivität (in %)		positiv prädiktiver Wert (in %)		Sensitivität (in %)		positiv prädiktiver Wert (in %)	
		Parameter	KI	Parameter	KI	Parameter	KI	Parameter	KI
Verfehlen des Regelstandards (der 4. Klasse) (BR = 17 %/14 %)	T_M (SR = 27 %/27 %)	69	[57;80]	43	[34;53]	61	[48;73]	33	[25;42]
	N_M (SR = 8 %/8 %)	33	[23;45]	67	[49;81]	35	[24;49]	61	[43;77]
	$T_M o N_M$ (SR = 29 %/29 %)	78	[66;87]	45	[36;54]	71	[58;82]	35	[27;44]
	$T_M u N_M$ (SR = 6 %/6 %)	25	[16;37]	69	[48;86]	26	[16;38]	62	[41;80]
Verfehlen der Noten 1, 2 oder 3(BR = 15 %/18 %)	T_M (SR = 27 %/27 %)	64	[51;76]	36	[27;45]	57	[45;68]	38	[29;48]
	N_M (SR = 8 %/8 %)	43	[30;56]	75	[58;88]	36	[26;48]	78	[61;90]
	$T_M o N_M$ (SR = 29 %/29 %)	75	[63;85]	38	[30;48]	66	[55;77]	41	[32;50]
	$T_M u N_M$ (SR = 6 %/6 %)	33	[22;46]	81	[61;93]	27	[18;39]	81	[61;93]

Anmerkungen. KI: 95%-Konfidenzintervall, BR: Basisrate, SR: Selektionsrate, T_M: Testergebnis in Mathematik, N_M: Schulnote in Mathematik, $T_M o N_M$: Testergebnis in Mathematik oder Schulnote in Mathematik, $T_M u N_M$: Testergebnis in Mathematik und Schulnote in Mathematik.

ESM 36

Details zur Bestimmung von Referenzwerten für die Klassifikationsgüte im Rahmen der Ergänzungsstudie

Um die Ergebnisse zur prognostischen Klassifikationsgüte von Bista-Tests zu bewerten, ist eine Einordnung in den bisherigen Forschungsstand von deutschsprachigen Schulleistungstests sehr hilfreich. Um diese Einordnung vorzunehmen ist es essentiell, dass zentrale Parameter, die Einfluss auf die Klassifikationsgüte nehmen, zwischen den verschiedenen Studien vergleichbar sind (s. a. den Abschnitt „Einordnung der Befunde in den Forschungsstand“ im Manuskript):

- Für Indizes zur Bewertung der Klassifikationsgüte wie bspw. Sensitivität (*Sen*) und positiv prädiktiver Wert (*ppW*) gilt, dass diese die „Leistungsfähigkeit eines Verfahrens unter bestimmten Anwendungsbedingungen [charakterisieren], die durch die Basis- und Selektionsraten bestimmt werden“ (Tröster, 2009, S. 139). Dies gilt insbesondere für den *ppW* (Petscher, Kim & Foorman, 2011; Streiner, 2003), aber auch für die *Sen* (Kilgus, Methe, Maggin & Tomasula, 2014; Whiting, Rutjes, Westwood & Mallett, 2013).
- Bei längeren Prognosezeiträumen ist von einer „validity degradation“ auszugehen: Längere Prognosezeiträume gehen in der Regel mit einer geringeren Prognosegenauigkeit einher (Dahlke, Kostal, Sackett & Kuncel, 2018; *Sen*: Kilgus et al., 2014).

Dies hatte mehrere Implikationen.

(1) Es gibt für den deutschsprachigen Raum zur Einordnung der Befunde der Hauptstudie keine Vorgängerstudien mit vergleichbarem Prognosezeitraum sowie vergleichbaren Basis- und Selektionsraten (s. Tabelle ESM-36.1). Eine Einordnung der Ergebnisse der Hauptstudie in den Forschungsstand bringt unseres Erachtens daher keinen zusätzlichen Erkenntnisgewinn, sondern ist eher irreführend. Für die Hauptstudie wurden daher auch keine Referenzwerte bestimmt, die auf Vorgängerstudien aufbauen.

Tabelle ESM-36.1

Gegenüberstellung von Studienparametern (Prognosezeitraum, Selektions- und Basisraten) der Hauptstudie (HS) und des Forschungsstandes (FS) deutscher Schulleistungstests

Mathematik	Anzahl der Studien-ergebnisse	Prognosezeitraum (25-75% Quantil)	Selektionsrate (25-75% Quantil)	Basisrate (25-75% Quantil)
Prognose von Testleistung				
FS: Sensitivität, Spezifität	**37**	**1-2**	**13-20**	**15-18**
FS: positiv/negativ prädiktiver Wert	**10**	**1-2**		
HS: SMBG		5	27 (57)	50
HS: Gymnasien		5	4 (16)	5
Prognose von Schulnote				
FS: Sensitivität, Spezifität	**4**	**1-1.5**	**16-22**	**14-15**
FS: positiv/negativ prädiktiver Wert	**3**	**1-1.3**		
HS: SMBG		5	15 (57)	32
HS: Gymnasien		5	4 (16)	26
Deutsch (i. w. S. Lesen)				
Prognose von Testleistung				
FS: Sensitivität, Spezifität	**19**	**1-1.3**	**12-17**	**20-21**
FS: positiv/negativ prädiktiver Wert	**5**	**0.5-3.0**		
HS: SMBG		5	15 (30)	18
HS: Gymnasien		5	1 (4)	1
Prognose von Schulnote				
FS: Sensitivität, Spezifität	**1**	**3**	**44**	**13**
FS: positiv/negativ prädiktiver Wert	**2**	**2.3-2.8**		
HS: SMBG		5	15 (30)	22
HS: Gymnasien		5	1 (4)	11

Anmerkungen. FS = Forschungsstand, HS = Hauptstudie, SMBG = Schulen mit mehreren Bildungsgängen. Werte in Klammern geben die Selektionsrate an mit Orientierung am Regelstandard.

(2) Für die Ergänzungsstudie gibt es Vorgängerstudien mit standardisierten Tests, die in zentralen Parametern (SR, BR, Prognosezeitraum) vergleichbar sind (s. Tabelle ESM-36.2). Für die Ergänzungsstudie haben wir folglich auch Referenzwerte für standardisierte Schulleistungstests im deutschsprachigen Raum bestimmt, die die Klassifikationsgüte abbilden, die man üblicherweise unter den Rahmenbedingungen der Ergänzungsstudie findet (s. Abbildung 3 im Artikel, sowie ESM-12, ESM-29. Die exakten Werte der Referenzbereiche berichten wir in der Tabelle ESM-36.3.

Tabelle ESM-36.2

Gegenüberstellung von Studienparametern (Prognosezeitraum, Selektions- und Basisraten) der Ergänzungsstudie (ES) und des Forschungsstandes(FS) deutscher Schulleistungstests

Mathematik	Anzahl der Studien-ergebnisse	Prognosezeitraum (25-75% Quantil)	Selektionsrate (25-75% Quantil)	Basisrate (25-75% Quantil)
Prognose von Testleistung				
FS: Sensitivität, Spezifität	**37**	**1-2**	**13-20**	**15-18**
FS: positiv/negativ prädiktiver Wert	**10**	**1-2**		
ES: Grundschule		1; 2	8 (27)	17
Prognose von Schulnote				
FS: Sensitivität, Spezifität	**4**	**1-1.5**	**16-22**	**14-15**
FS: positiv/negativ prädiktiver Wert	**3**	**1-1.3**		
ES: Grundschule		1; 2	8 (27)	15

Anmerkungen. FS = Forschungsstand, ES = Ergänzungsstudie. Werte in Klammern geben die Selektionsrate an mit Orientierung am Regelstandard.

Tabelle ESM-36.3

Referenzwerte für die Ergänzungsstudie auf der Basis des Forschungsstandes für Schulleistungstests

	A) 4. zur 5. Klasse		B) 4. zur 6. Klasse	
Mathematik	25 % Q	75 % Q	25 % Q	75 % Q
Prognose von Testleistung				
Sensitivität	39	67	39	67
Spezifität	88	93	88	93
positiv prädiktiver Wert	52	57	45	50
negativ prädiktiver Wert	89	90	92	92
Prognose von Schulnote				
Sensitivität	37	46	37	46
Spezifität	90	94	90	94
positiv prädiktiver Wert	49	51	55	57
negativ prädiktiver Wert	91	91	88	89

Anmerkungen. Q = Quantil.

Wie haben wir die Referenzwerte bestimmt? Die Referenzwerte für die *Sen* (bzw. *Spe*) für die Ergänzungsstudie ermittelten wir auf der Basis aller relevanten Studien, die wir im ESM-1 dieses Artikels zum Forschungsstand zusammenfassen. Um relevante Studien zu identifizieren, wendeten wir die folgenden Kriterien an: (1) Um eine Konfundierung der Ergebnisse zu vermeiden, schlossen wir bei der Berechnung der Referenzwerte alle Studien

aus, in denen bildungsstandardbasierte Tests verwendet wurden. (2) Für den KR 3-4 wurde lediglich der Kennwert einbezogen, der für eine Selektionsquote von 15 % ermittelt wurde. Einen Überblick über die einbezogenen Studienergebnisse, die für die Bestimmung der Referenzwerte von *Sen* und *Spe* herangezogen wurden, gibt ESM-36.4.

Tabelle ESM-36.4 (weiterführende Informationen in der Tabelle ESM-1)

Forschungsstand zur prognostischen Klassifikationsgüte standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	Sen (in %)	Spe (in %)	Quelle
Vorhersage von Mathematiktestergebnis auf Mathematiktestergebnis – kommerziell erhältliche Tests								
DEMAT 1 + (PR < 15)	3	1. → 4.	DEMAT 4 (PR < 15)	< 15	< 15	38	92	1
	2	1. → 3.	VERA 3 (Vorversion) (PR < 15)	15[b]	10[b]	28	94	
	2	1. → 3.	DEMAT 3 + (PR < 15)	< 15	< 15	39	93	
	1	1. → 2.	DEMAT 2 + (PR < 15)	< 15	< 15	44	95	
DEMAT 2 + (PR < 15)	2	2. → 4.	DEMAT 4 (PR < 15)	< 15	< 15	72	91	1
	1	2. → 3.	VERA 3 (Vorversion) (PR < 15)	15[b]	13[b]	45	93	
	1	2. → 3.	DEMAT 3 + (PR < 15)	< 15	< 15	56	91	
DEMAT 3 + (PR < 15)	1	3. → 4.	DEMAT 4 (PR < 15)	< 15	< 15	52	95	1
DEZ (n.e.)	≈ 1	0. → 1.	DEMAT 1 + (PR < 15)	6[b]	10[b]	100	96	2
ERT 0 + (PR < 16)	≈ 2	1. → 2.	DEMAT 2 + (PR < 16)	18	12	31	92	3
	≈ 1	1. → 1.	DEMAT 1 + (PR < 16)	16	15	39	89	
HaReT 1 (= HRT 1) (PR < 16)	≈ 2	1. → 2.	DEMAT 2 + (PR < 16)	19	13	35	92	3
	≈ 1	1. → 1.	DEMAT 1 + (PR < 16)	16	16	49	91	
Kalkulie 1 Diagnoseteil 1 (PR < 16)	≈ 2	1. → 2.	DEMAT 2 + (PR < 16)	19	11	29	93	3
	≈ 1	1. → 1.	DEMAT 1 + (PR < 16)	16	16	52	91	
KR 3-4	1	3. → 4.	DEMAT 3 + und 4 (PR < 25)	< 25	< 15	88	89	4
UGT (n.e.)	≈ 2	1. → 2.	AST 2 / Mathematik (t < 44)	13	10	36	94	5

(Tabelle ESM-36.4 wird fortgesetzt)

Test (SW)[a]	Dauer	Klassen	Kriterium (SW)[a]	BR (in %)	SR (in %)	Sen (in %)	Spe (in %)	Quelle
Vorhersage von Mathematiktestergebnis auf Mathematiktestergebnis – nicht kommerziell erhältliche Tests								
Testbatterie Dornheim (n.e.)	≈ 2.5	0. → 2.	DEMAT 2 + (PR < 16)	16[b]	25[b]	60	82	6
	+P (3)			16[b]	28[b]	70	80	
	≈ 2			16[b]	20[b]	60	88	
	+P (3)			16[b]	32[b]	80	77	
	≈ 1.5	0. → 1.	DEMAT 1 + (PR < 18)	18[b]	23[b]	67	86	
	+P (3)			18[b]	29[b]	79	82	
	≈ 1			18[b]	23[b]	67	86	
	+P (3)			18[b]	30[b]	79	81	
Testbatterie Kaufmann (n.e.)	≈ 2	1. → 2.	AST 2 / Mathematik (t < 44)	15	11	38	93	5
			Gruppen- und Einzeltest U3 (k.A.)	14	12	36	92	
M. L. U2 (Einzel- und Gruppentest) (n.e.)	≈ 1	1. → 1.	AST 2 / Mathematik (t < 44)	14	22	67	86	5
RT U2 (n.e.)	≈ 1	1. → 1.	AST 2 / Mathematik (t < 44)	14	12	36	92	5
Testbatterie Krajewski (n.e.)	≈ 2.5	0. → 2.	DEMAT 2 + (Vorversion) (PR < 15)	18[b]	11[b]	47	97	7
	+P(G)			18[b]	16[b]	53	92	
	≈ 1			14[b]	16[b]	47	90	
	+P(G)			14[b]	18[b]	53	88	
	≈ 1.5	0. → 1.	DEMAT 1 + (PR < 15)	15[b]	14[b]	61	94	
	+P(G)			15[b]	18[b]	61	89	
	≈ 1			14[b]	18[b]	65	90	
	+P(G)			14[b]	20[b]	65	87	
Vorhersage von Mathematiktestergebnis auf Mathematiknote – kommerziell erhältliche Tests								
UGT (n.e.)	1	2. → 2.	Note (Note > 3)	19[b]	9	18	93	5
Vorhersage von Mathematiktestergebnis auf Mathematiknote – nicht kommerziell erhältliche Tests								
M. L. U2 (n.e.)	1	2. → 2.	Note (Note > 3)	20	21	46	85	5
RT U2 (n.e.)	1	2. → 2.	Note (Note > 3)	21	16	43	91	5
Testbatterie Kaufmann (n.e.)	≈ 2	1. → 2.	Note (Note > 3)	21	11	48	99	5

Anmerkungen. SW = Schwellenwert, Dauer = Länge des Vorhersagezeitraums in Jahren, Klassen = Klassen des ersten und zweiten Messzeitpunktes (Vorhersagezeitraum), BR = Basisrate, SR = Selektionsrate, Sen = Sensitivität, Spe = Spezifität, DEMAT = Deutscher Mathematiktest – für erste, zweite, dritte und vierte Klassen, VERA 3 = Vergleichsarbeiten in der dritten Klassenstufe, DEZ = Diagnostikum zur Entwicklung des Zahlkonzepts, ERT 0 + = Eggenberger Rechentest, HaReT 1 = Hamburger Rechentest für Klasse 1, Kalkulie 1 = Kalkulie Diagnoseaufgaben Teil 1, KR 3-4 = Kettenrechner für 3. und 4. Klassen, UGT = Utrechter Zahlbegriffstest, AST 2 / Mathematik = Allgemeiner Schulleistungstest für 2. Klassen, +P = Zusätzlich wurde zum Testergebnis ein weiteres Testergebnis zur Prognose herangezogen: (3) = Zahlen Lesen, sprachliche Arbeitsgedächtnisleistung, räumliche IQ-Komponente (2 Risikopunkte), M. L. U2 = Mathematischer Leistungstest U2, RT U2 = Rechentest U2, (G) = Gedächtniskapazität.

[a] Schwellenwert (SW) Abkürzungen: PR = Prozentrang, n.e. = Schwellenwert nicht eindeutig angegeben, t = t-Wert, k.A. = keine Angabe, Note = Schulnote, TW = Testwert, KS = Kompetenzstufe.

[b] selbst berechnete Werte (Tröster, 2009, S. 105).

Literatur der Tabelle ESM-36.4

1 Hasselhorn, M., Roick, T. & Gölitz, D. (2005). Stabilitäten und prognostische Validitäten der Mathematikleistungen. Eine Längsschnittstudie mit der DEMAT-Reihe in der Grundschule (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, H. Marx & W. Schneider (Hrsg.), *Diagnostik von Mathematikleistungen* (Band 4, S. 187–198). Göttingen: Hogrefe.

2 Weißhaupt, S., Peucker, S. & Wirtz, M. (2006). Diagnose mathematischen Vorwissens im Vorschulalter und Vorhersage von Rechenleistungen und Rechenschwierigkeiten in der Grundschule. *Psychologie in Erziehung und Unterricht: Zeitschrift für Forschung und Praxis*, (4), 236–245.

3 Gomm, B. (2014). *Prognostische Validität mathematischer Screenings*. Dortmund: Technische Universität Dortmund. Verfügbar unter: https://eldorado.tu-dortmund.de/bitstream/2003/33789/1/Dissertation.pdf

4 Gölitz, D., Roick, T. & Hasselhorn, M. (2013). Kettenrechner für dritte und vierte Klassen (KR 3-4) (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, A. Heinze, W. Schneider & U. Trautwein (Hrsg.), *Diagnostik mathematischer Kompetenzen* (Band 11, S. 149–164). Göttingen: Hogrefe.

5 Kaufmann, S. (2002). *Früherkennung von Rechenstörungen in der Eingangsklasse der Grundschule und darauf abgestimmte remediale Massnahmen* (Europäische Hochschulschriften. Reihe XI, Pädagogik). Frankfurt am Main: Peter Lang.

6 Dornheim, D. (2007). *Prädiktion von Rechenleistung und Rechenschwäche: Der Beitrag von Zahlen-Vorwissen und allgemein-kognitiven Fähigkeiten*. Berlin: Logos Verlag Berlin GmbH.

7 Krajewski, K. (2003). *Vorhersage von Rechenschwäche in der Grundschule* (2. Auflage). Hamburg: Dr. Kovač.

Auf Basis der ausgewählten Studien berechneten wir dann die Werte für die Sensitivität (*Sen*) bzw. Spezifität (*Spe*) separat für jede der vier möglichen Prädiktor-Kriterium-Kombinationen unserer Studie (Mathematik: Prädiktor Testergebnis und Kriterium Testergebnis sowie Prädiktor Testergebnis und Kriterium Note; Deutsch: Prädiktor Testergebnis und Kriterium Testergebnis; Prädiktor Testergebnis und Kriterium Note). Als Referenzwerte nutzten wir

dabei jeweils das 25 % Quantil und das 75 % Quantil der *Sen* bzw. *Spe* der relevanten Vorgängerstudien, die in Tabelle ESM-36.4 dargestellt werden.

Die Referenzwerte für die positiv prädiktiven Werte (*ppW*) (bzw. negativ prädiktiven Werte, *npW*) haben wir nach der Methode von Taylor und Russel (1939) mit dem R-Paket iopsych (Goebl, Jones & Beatty, 2016) unter Nutzung der Funktion trmodel geschätzt[7]. Dies hat den Vorteil, dass auf diese Weise Referenzwerte der prädiktiven Werte auf der Basis vergleichbarer Anwendungsbedingungen geschätzt werden können. Die Schätzung der Referenzwerte erfolgte auf der Grundlage der Basisraten unserer Ergänzungsstudie und einer festgelegten typischen Selektionsrate von 15 % (s. Artikel Abschnitt „Forschungsstand zur prognostischen Klassifikationsgüte standardisierter Schulleistungstests und von Schulnoten“ und ESM-1). Die prognostischen Korrelationskoeffizienten basieren auf dem Forschungsstand, den Fuchs und Brunner (2017, ESM-1) berichten. Relevante Vorgängerstudien wurden unter Anwendung der folgenden Kriterien ausgewählt: (1) Um eine Konfundierung der Ergebnisse zu vermeiden, schlossen wir alle Studien aus, in denen bildungsstandardbasierte Tests verwendet wurden. (2) Zudem wurden lediglich die berichteten Korrelationskoeffizienten auf der Grundlage bivariater Analysen einbezogen; adjustierte Koeffizienten zur Bestimmung des prognostischen Mehrwerts wurden nicht berücksichtigt. Die einbezogenen Studienergebnisse sind in Tabelle ESM-36.5 dargestellt.

Auf Basis der ausgewählten Studien bestimmten wir dann die Referenzwerte, in dem wir jeweils das 25 % Quantil und das 75 % Quantil für die geschätzten *ppW* bzw. geschätzten *npW* berechneten und zwar für jede der vier möglichen Prädiktor-Kriterium-Kombinationen.

[7] Die Methode von Taylor und Russel (1939) ist ein etabliertes Verfahren zur Abschätzung des positiv prädiktiven Wertes. So haben wir exemplarisch für die Ergänzungsstudie für den Mathematiktest festgestellt, dass die geschätzten prädiktiven Werte auf der Grundlage der Methode von Taylor und Russel (1939) sehr nah an unseren empirisch ermittelten Werten liegen: der Median der Differenz für den *ppW* ($ppW_{\text{Taylor \& Russel}}$ - $ppW_{\text{empirisch}}$) lag bei -1 %, (Min. = -10 %, Max. = 7 %); der Median der Differenz für den *npW* ($npW_{\text{Taylor \& Russel}}$ - $npW_{\text{empirisch}}$) lag bei 0 % (Min. = -1 %, Max. = 2 %).

Tabelle ESM-36.5 (weiterführende Informationen in Fuchs & Brunner, 2017, Tabelle ESM-1)

Forschungsstand zur prognostischen Klassifikationsgüte standardisierter Schulleistungstests in Bezug auf spätere Testleistungen und Schulnoten

Test	Dauer	Klassen	Kriterium	Korrelation	Quelle
Zusammenhang zwischen Mathematiktest und Mathematiktest					
DEMAT 1+	3	1.- 4.	DEMAT 4	r = .59	1
DEMAT 1+	2	1.- 3.	DEMAT 3+	r = .55	1
DEMAT 2+	2	2.- 4.	DEMAT 4	r = .64	1
DEMAT 1+	1	1.- 2.	DEMAT 2+	r = .67	1
DEMAT 2+	1	2.- 3.	DEMAT 3+	r = .65	1
DEMAT 1+	1	1.- 2.	DEMAT 2+	r = .80	2
DEMAT 3+	1	3.- 4.	DEMAT 4	r = .68	3
MBK-1	2	1.- 2.	HRT 1-4	r = .71	4
MBK-1	2	1.- 2.	HRT 1-4	r = .64	4
MBK-1	1	1.- 2.	DEMAT 1+	r = .72	4
Zusammenhang zwischen Mathematiktest und Mathematiknote					
DEMAT 2+	2	2.- 4.	Note	r = -.64	5
DEMAT 2+	1	2.- 3.	Note	r = -.67	5
DEMAT 3+	1	3.- 4.	Note	r = -.69	3

Anmerkungen. Dauer = Länge des Vorhersagezeitraums in Jahren, Klassen = Klassen des ersten und zweiten Messzeitpunktes (Vorhersagezeitraum), DEMAT = Deutscher Mathematiktest – für erste, zweite, dritte und vierte Klassen, MBK-1 = Test zur Erfassung mathematischer Basiskompetenzen, HRT 1-4 = Heidelberger Rechentest. Aus „Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? von G. Fuchs & M. Brunner, 2017, *Zeitschrift für Pädagogische Psychologie, 31* (1), ESM-1, S. 1 und S. 3. Copyright 2017 bei Hogrefe. Veränderte Wiedergabe.

Literatur der Tabelle ESM-36.5

1 Hasselhorn, M., Roick, T. & Gölitz, D. (2005). Stabilitäten und prognostische Validitäten der Mathematikleistungen. Eine Längsschnittstudie mit der DEMAT-Reihe in der Grundschule (Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik). In M. Hasselhorn, H. Marx & W. Schneider (Hrsg.), *Diagnostik von Mathematikleistungen* (Band 4, S. 187–198). Göttingen: Hogrefe.

2 Dornheim, D. (2007). *Prädiktion von Rechenleistung und Rechenschwäche. Der Beitrag von Zahlen-Vorwissen und allgemein-kognitiven Fähigkeiten*. Berlin: Logos Verlag Berlin GmbH.

3 Roick, T., Gölitz, D. & Hasselhorn, M. (2004). *DEMAT 3+: Deutscher Mathematiktest für dritte Klassen*. Göttingen: Beltz.

4 Sinner, D., Ennemoser, M. & Krajewski, K. (2011). Entwicklungspsychologische Frühdiagnostik mathematischer Basiskompetenzen im Kindergarten- und frühen Grundschulalter (MBK-0 und MBK-1). In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen* (Band 9, S. 109–126). Göttingen: Hogrefe Verlag GmbH & Co. KG.

5 Krajewski, K., Liehm, S. & Schneider, W. (2004). *DEMAT 2+: Deutscher Mathematiktest für zweite Klassen*. Göttingen: Beltz.

ESM-36: Literatur

Dahlke, J. A., Kostal, J. W., Sackett, P. R. & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. Advance online publication. *Journal of Applied Psychology*. doi:10.1037/apl0000316

Fuchs, G. & Brunner, M. (2017). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie, 31* (1), 27–39. doi:10.1024/1010-0652/a000195

Goebl, A. P., Jones, J. A. & Beatty, A. S. (2016). iopsych: Methods for industrial/organizational psychology. (Version 0.90).

Kilgus, S. P., Methe, S. A., Maggin, D. M. & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52* (4), 377–405. doi:10.1016/j.jsp.2014.06.002

Petscher, Y., Kim, Y.-S. & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention, 36* (3), 158–166. doi:10.1177/1534508410396698

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81* (3), 209–219. doi:10.1207/S15327752JPA8103_03

Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23* (5), 565–578. doi:10.1037/h0057079

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter: Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E. & Mallett, S. (2013). A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology, 66* (10), 1093–1104. doi:10.1016/j.jclinepi.2013.05.014

DANKSAGUNG

Die Danksagung ist in der Online-Version
aus Gründen des Datenschutzes nicht enthalten.

ERKLÄRUNG

Hiermit versichere ich, dass ich die Dissertation „Prognosegüte bildungsstandardbasierter Tests“ selbständig verfasst habe. Alle Hilfsmittel, die ich verwendet habe, sind angegeben. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder angelehnt worden.

Potsdam, im Dezember 2018

Gesine Fuchs

EIGENANTEIL UND VERÖFFETNLICHUNGEN

Die folgende Tabelle veranschaulicht den Eigenanteil an den veröffentlichten oder zur Veröffentlichung eingereichten wissenschaftlichen Schriften innerhalb meiner Dissertationsschrift.

Autoren	Titel	Status	Eigenanteil
Fuchs, G. & Brunner, M.	Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen?	2017 Veröffentlicht in *Zeitschrift für Pädagogische Psychologie*, *31* (1), 27-39. https://doi.org/10.1024/1010-0652/a000195	Konzeption der Fragestellung (überwiegend), Aufarbeitung der Literatur und des theoretischen Hintergrunds (vollständig); Datenaufbereitung (überw. ab Datenübergabe), Datenanalyse (vollst.); Verfassung des Manuskriptes (überw.), Antworten auf Gutachten (überw.)
Fuchs, G., Nachtigall, C. & Brunner, M.	Wie stark variiert die Güte bildungsstandard-basierter Tests zur Prognose zukünftiger schulischer Leistungen zwischen Schulen? – Ein metanalytischer Ansatz zur Validitäts-generalisierung	Eingereicht	Konzeption der Fragestellung (überw.), Aufarbeitung der Literaturund des theore-tischen Hintergrunds (vollst.); Datenaufbereitung (vollst. ab Datenübergabe), Datenanalyse (vollst.); Verfassung des Manuskriptes (überw.)
Fuchs, G., Nachtigall, C., Harych, P. & Brunner, M.	Wie gut lassen sich mit bildungsstandard-basierten Kompetenztests Kinder identifizieren, die wichtige Bildungs-ergebnisse im Verlauf der Schulkarriere verfehlen? Ergebnisse zweier Längsschnitt-studien zur Klassifika-tionsgüte in den Fächern Mathematik und Deutsch.	Überarbeitetes Manuskript in Begutachtung	Konzeption der Fragestellung (überw.), Aufarbeitung der Literatur und des theoretischen Hintergrunds (vollst.); Datenaufbereitung (überw. ab Datenübergabe), Datenanalyse (vollst.); Verfassung des Manuskriptes (überw.), Antworten auf Gutachten (überw.)

Potsdam, im Dezember 2018

Gesine Fuchs

LEBENSLAUF

Der Lebenslauf ist in der Online-Version

aus Gründen des Datenschutzes nicht enthalten.