Leistungsstarke Schülerinnen und Schüler in Deutschland:

Definition, Entwicklung, soziale Integration

Dissertation
zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.) im Fach Erziehungswissenschaften

eingereicht bei der
Humanwissenschaftlichen Fakultät der
Universität Potsdam

vorgelegt von
Claudia Neuendorf, M.Sc.

2022

Verteidigt am 14. Oktober 2022 in Potsdam

Gutachterinnen

Prof. Dr. Miriam Vock, Universität Potsdam
Prof. Dr. Petra Stanat, Humboldt-Universität zu Berlin

Online veröffentlicht auf dem
Publikationsserver der Universität Potsdam:
https://doi.org/10.25932/publishup-56470
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-564702

Zusammenfassung

Die vorliegende kumulative Promotionsarbeit beschäftigt sich mit leistungsstarken Schülerinnen und Schülern, die seit 2015 in der deutschen Bildungspolitik, zum Beispiel im Rahmen von Förderprogrammen wieder mehr Raum einnehmen, nachdem in Folge des „PISA-Schocks“ im Jahr 2000 zunächst der Fokus stärker auf den Risikogruppen lag. Während leistungsstärkere Schülerinnen und Schüler in der öffentlichen Wahrnehmung häufig mit „(Hoch-)Begabten“ identifiziert werden, geht die Arbeit über die traditionelle Begabungsforschung, die eine generelle Intelligenz als Grundlage für Leistungsfähigkeit von Schülerinnen und Schülern begreift und beforscht, hinaus. Stattdessen lässt sich eher in den Bereich der Talentforschung einordnen, die den Fokus weg von allgemeinen Begabungen auf spezifische Prädiktoren und Outcomes im individuellen Entwicklungsverlauf legt. Der Fokus der Arbeit liegt daher nicht auf Intelligenz als Potenzial, sondern auf der aktuellen schulischen Leistung, die als Ergebnis und Ausgangspunkt von Entwicklungsprozessen in einer Leistungsdomäne doppelte Bedeutung erhält.

Die Arbeit erkennt die Vielgestaltigkeit des Leistungsbegriffs an und ist bestrebt, neue Anlässe zu schaffen, über den Leistungsbegriff und seine Operationalisierung in der Forschung zu diskutieren. Hierfür wird im ersten Teil ein systematisches Review zur Operationalisierung von Leistungsstärke durchgeführt (Artikel I). Es werden Faktoren herausgearbeitet, auf welchen sich die Operationalisierungen unterscheiden können. Weiterhin wird ein Überblick gegeben, wie Studien zu Leistungsstarken sich seit dem Jahr 2000 auf diesen Dimensionen verorten lassen. Es zeigt sich, dass eindeutige Konventionen zur Definition schulischer Leistungsstärke noch nicht existieren, woraus folgt, dass Ergebnisse aus Studien, die sich mit leistungsstarken Schülerinnen und Schülern beschäftigen, nur bedingt miteinander vergleichbar sind. Im zweiten Teil der Arbeit wird im Rahmen zwei weiterer Artikel, welche sich mit der Leistungsentwicklung (Artikel II) und der sozialen Einbindung (Artikel III) von leistungsstarken Schülerinnen und Schülern befassen, darauf aufbauend der Ansatz verfolgt, die Variabilität von Ergebnissen über verschiedene Operationalisierungen von Leistungsstärke deutlich zu machen. Damit wird unter anderem auch die künftige Vergleichbarkeit mit anderen Studien erleichtert. Genutzt wird dabei das Konzept der Multiversumsanalyse (Steegen et al., 2016), bei welcher viele parallele Spezifikationen, die zugleich sinnvolle Alternativen für die Operationalisierung darstellen, nebeneinandergestellt und in ihrem Effekt verglichen werden (Jansen et al., 2021). Die Multiversumsanalyse knüpft konzeptuell an das bereits vor längerem entwickelte Forschungsprogramm des kritischen Multiplismus an (Patry, 2013; Shadish, 1986, 1993), erhält aber als spezifische Methode aktuell

im Rahmen der Replizierbarkeitskrise in der Psychologie eine besondere Bedeutung. Dabei stützt sich die vorliegende Arbeit auf die Sekundäranalyse großangelegter Schulleistungsstudien, welche den Vorteil besitzen, dass eine große Zahl an Datenpunkten (Variablen und Personen) zur Verfügung steht, um Effekte unterschiedlicher Operationalisierungen zu vergleichen.

Inhaltlich greifen Artikel II und III Themen auf, die in der wissenschaftlichen und gesellschaftlichen Diskussion zu Leistungsstarken und ihrer Wahrnehmung in der Öffentlichkeit immer wieder aufscheinen: In Artikel II wird zunächst die Frage gestellt, ob Leistungsstarke bereits im aktuellen Regelunterricht einen kumulativen Vorteil gegenüber ihren weniger leistungsstarken Mitschülerinnen und Mitschülern haben (Matthäus-Effekt). Die Ergebnisse zeigen, dass an Gymnasien keineswegs von sich vergrößernden Unterschieden gesprochen werden kann. Im Gegenteil, es verringerte sich im Laufe der Sekundarstufe der Abstand zwischen den Gruppen, indem die Lernraten bei leistungsschwächeren Schülerinnen und Schülern höher waren. Artikel III hingegen betrifft die soziale Wahrnehmung von leistungsstarken Schülerinnen und Schülern. Auch hier hält sich in der öffentlichen Diskussion die Annahme, dass höhere Leistungen mit Nachteilen in der sozialen Integration einhergehen könnten, was sich auch in Studien widerspiegelt, die sich mit Geschlechterstereotypen Jugendlicher in Bezug auf Schulleistung beschäftigen. In Artikel III wird unter anderem erneut das Potenzial der Multiversumsanalyse genutzt, um die Variation des Zusammenhangs über Operationalisierungen von Leistungsstärke zu beschreiben. Es zeigt sich unter unterschiedlichen Operationalisierungen von Leistungsstärke und über verschiedene Facetten sozialer Integration hinweg, dass die Zusammenhänge zwischen Leistung und sozialer Integration insgesamt leicht positiv ausfallen. Annahmen, die auf differenzielle Effekte für Jungen und Mädchen oder für unterschiedliche Fächer abzielen, finden in diesen Analysen keine Bestätigung.

Die Dissertation zeigt, dass der Vergleich unterschiedlicher Ansätze zur Operationalisierung von Leistungsstärke — eingesetzt im Rahmen eines kritischen Multiplismus — das Verständnis von Phänomenen vertiefen kann und auch das Potenzial hat, Theorieentwicklung voranzubringen.

Inhalt

1

Einleitung und theoretischer Rahmen

1. Einleitung und theoretischer Rahmen

1.1. Einleitung und Struktur der Dissertation

Zwei der zentralen Funktionen des Bildungssystems sind die Vermittlung von Kenntnissen und Fähigkeiten sowie die Verteilung von gesellschaftlichen Chancen und Positionen nach dem Leistungsprinzip als allgemein anerkanntem Maßstab (Fend, 2008). Gelingt es der Institution Schule, diese Funktionen zu erfüllen, dann stellt sie die Produktivität des Wirtschaftssystems und die Reproduktion und Funktionsfähigkeit des Sozialsystems auf der einen Seite sicher und ermöglicht auf der anderen Seite den Individuen, die sie durchlaufen haben, an der Gesellschaft teilzuhaben, ihre individuellen Leistungspotenziale zu entwickeln und ihre berufliche und gesellschaftliche Stellung durch eigene Lernanstrengungen und Erbringung schulischer Leistungen in die Hand zu nehmen. Schulleistung wird so zu einem der wichtigsten „Produkte" des Schulsystems, was sich darin äußert, dass die Schuleffektivitätsforschung die von Schülerinnen und Schülern erbrachte Leistung als Hauptkriterium der „effektiven" Schule untersucht.

Der an diesem Kriterium gemessene Erfolg des Bildungssystems stellte sich nach den Ergebnissen der PISA-Studie im Jahr 2000 enttäuschend dar. 10 % der Schülerschaft, so die Aussage, fehlten die basalen Grundkompetenzen, um an der Gesellschaft teilhaben zu können (Baumert et al., 2002; Organisation for Economic Co-operation and Development [OECD], 2001). Angesichts dieser Resultate stellte sich die Förderung von Schülerinnen und Schülern am unteren Ende der Leistungsverteilung als dringlichstes Problem dar, welches in der Folge mit Nachdruck angegangen wurde (z. B. im Rahmen der *Förderstrategie für leistungsschwächere Schülerinnen und Schüler*; Kultusministerkonferenz [KMK], 2010). Für die Funktions- und Wettbewerbsfähigkeit von Wirtschaft und Gesellschaft sind jedoch nicht nur die Basiskompetenzen entscheidend, sondern auch der Anteil von Menschen, die in der Lage sind, Höchstleistungen in unterschiedlichen Bereichen zu erbringen.

Akademisch leistungsstarke Schülerinnen und Schüler sind daher ab dem Jahr 2015 wieder in den Fokus der Bildungspolitik und, vermittelt über entsprechende Forschungsförderung, auch stärker in den Fokus der Bildungsforschung geraten. Dies hat die vorliegende Dissertation zum Anlass genommen, sich mit dem Thema akademisch leistungsstarker Schülerinnen und Schüler auseinanderzusetzen. Dabei ist schnell deutlich geworden, dass eine begriffliche Unschärfe existiert, was genau mit Leistungsstärke gemeint ist. Aus diesem Grund wird in den folgenden Kapiteln zunächst ein kurzer Abriss vorgenommen, in welchen Forschungsbereichen Leistung bei Schülerinnen und Schülern typischerweise untersucht wird und was sich daraus für die Eingrenzung des Gegenstandes ableiten lässt. Daraufhin wird dargestellt, wie leistungsstarke

Schülerinnen und Schüler bisher für verschiedene Zwecke, nämlich in der Bildungsberichterstattung, in der pädagogischen Praxis und in der Forschung identifiziert werden.

Der erste der drei Beiträge dieser Dissertation, der sich mit der Operationalisierung von Leistungsstärke in der Forschung beschäftigt, wird daraufhin vorgestellt und diskutiert. Aus dieser Diskussion heraus wird das Vorgehen der weiteren Arbeit abgeleitet. Da die beiden weiteren Beiträge jeweils die Methode auf ein unterschiedliches inhaltliches Thema anwenden, wird jedem Beitrag ein eigener Theorieteil gewidmet. Eine Gesamtdiskussion am Ende der Arbeit führt die Ergebnisse dann übergreifend zusammen.

1.2. Theoretische Konzepte von Leistung und Leistungsstärke

Leistung und leistungsstarke Schülerinnen und Schüler werden in unterschiedlichen Teilbereichen der Pädagogik und Psychologie untersucht. Die für die Untersuchung von Hochleistungen relevantesten Forschungsansätze sind die Begabungs- und Expertiseforschung, welche spezifisch auf Menschen mit besonderem Leistungspotenzial bzw. Personen, die besondere Leistungen erbringen, fokussieren. Daneben existieren in unterschiedlichen Bereichen der Psychologie, in der Erziehungswissenschaft und Soziologie Begriffe von Leistung, aus denen sich Folgen für die Betrachtung akademisch Leistungsstarker ergeben. Diese unterschiedlichen Perspektiven werden hier im Folgenden zunächst vorgestellt.

1.2.1. Begabungs- und Expertiseforschung

Die psychologische Begabungsforschung hat ihre Wurzeln zu Beginn des 20. Jahrhunderts und entstand zeitgleich mit der Entwicklung der ersten Intelligenztests. Sie bezog sich entsprechend auf die Untersuchung und Beschreibung von Personen, die Höchstleistungen in kognitiven Grundfähigkeiten, wie sie typischerweise von Intelligenztests gemessen werden, erbringen. Als Väter der Begabungsforschung werden meist Alfred Binet und Theodore Simon, Entwickler des ersten Intelligenztests im Jahr 1904 (Schweizer, 2006), und Lewis Terman benannt, der mit seiner 1921 begonnenen längsschnittlichen Untersuchung Hochbegabter den Grundstein für die Begabungsforschung legte (Terman, 1925). Mit seiner Studie verfolgte er das Ziel, etwa 1500 Kinder, die im Rahmen von IQ-Tests als „Genies“ identifiziert wurden (weniger als ein Prozent der Population) zu charakterisieren und ihre Entwicklung über die Zeit zu verfolgen (Terman, 1954). Terman machte bereits in den Anfängen der Begabungsforschung einen Unterschied zwischen der angeborenen breit angelegten intellektuellen Leistungsfähigkeit auf der einen Seite (Begabung / *giftedness / intelligence*), den domänenspezifischen Talenten (*talents*) und den Leistungen (*achievement* / *accomplishments*), die auf dieser Grundlage erfolgen können, für deren Ausbildung allerdings noch weitere Faktoren eine Rolle spielen (Terman, 1954). Folglich wurden

bald Modelle entwickelt, die erklären sollten, wie Leistungsexzellenz zustande kommt, wobei der Begabung als notwendiger Bedingung für die Entwicklung herausragender Leistungen zunächst eine Hauptrolle zukam. Während die Definition und Messung von Begabung, also Leistungspotenzial, Gegenstand wissenschaftlicher Diskussion, Modelle und Kontroversen war, wurde der Leistung vergleichsweise wenig Aufmerksamkeit gezollt und im weiteren Verlauf teilweise auch häufig nicht konzeptuell von Begabung abgegrenzt (Harder, Vialle & Ziegler, 2014).

Immerhin fand die Leistung mit der Entwicklung von Moderatorenmodellen (z. B. Gagné, 1985; Heller, 2001) expliziten Eingang in die Modellbildung in der Begabungsforschung. In diesen Modellen werden die Bedingungen der Transformation von Potenzial als Prädiktor in Leistung bzw. Performanz als Kriterium betrachtet (Gagné, 1985; Hany, 2012; Harder, 2012; Harder et al., 2014). Solche Modelle enthalten meist eine Anzahl von Begabungsfaktoren (z.B. Intelligenz, Kreativität, Musikalität), die mit nicht-kognitiven Persönlichkeitsmerkmalen (z.B. Motivationale und emotionale Dispositionen, Selbstregulation) und Umweltmerkmalen (z. B. familiärer Hintergrund, Lernumwelt) interagieren, um schließlich eine Leistung in einem spezifischen Leistungsbereich hervorzubringen (Heller, 2001).

Die Rückführung von außergewöhnlichen Leistungen auf eine hohe Begabung weckte jedoch auch Widerspruch: Da für besondere Leistungen besonders hohe kognitive Fähigkeiten nicht unbedingt ausschlaggebend seien, sondern auch Personen mit durchschnittlichen kognitiven Fähigkeiten Expertise in einem Bereich entwickeln und bedeutsame Leistungen erbringen könnten (Ericsson, Roring & Nandagopal, 2007), entwickelte sich in den 1990er Jahren die Expertiseforschung.

Bei diesem Paradigma sollte der Vergleich von Informationsverarbeitungsprozessen zwischen Experten und Novizen in einer Domäne Aufschluss über das Wesen von Expertise geben (Ericsson & Charness, 1994). Dabei wurde die Bedeutung von Übungsprozessen für die Genese von Expertise (als Voraussetzung für Leistungsexzellenz) in den Mittelpunkt gestellt. Der Begriff der Expertise ist jedoch einer, der sich nur bedingt auf Schülerinnen und Schüler anwenden lässt, da die langjährige Übungspraxis (*deliberate practice*), die für die Entwicklung von Expertise in einem Gebiet vorausgesetzt wird, in diesem Alter meist noch nicht erfolgt sein kann. Darüber hinaus werden typischerweise nicht schulische Leistungsdomänen untersucht, sondern spezifische Fertigkeiten wie Schachspielen, Musizieren, Leistungssport oder berufliche Tätigkeiten. Dennoch spielt die Expertiseforschung eine wichtige Rolle für die Weiterentwicklung der schulisch orientierten Begabungsforschung, denn sie verlagerte den Fokus noch stärker auf den dynamischen Aspekt der Potenzialentwicklung.

Die neueren dynamischen Talententwicklungsmodelle stellen eine Verbindung zwischen Begabungs- und Expertiseforschung dar (Perleth, 2001; Preckel et al., 2020; Subotnik, Olszewski-Kubilius & Worrell, 2011). Bei ihnen ist die Unterscheidung zwischen Begabung als Potenzial auf der einen und Leistung als Ergebnis auf der anderen Seite ein Stück weit wieder aufgehoben, indem der Entwicklungsprozess von allgemeinen zu immer spezifischeren Kompetenzen und Fähigkeiten beschrieben und stärker als ein Kontinuum verstanden wird. Die Expertiseforschung wird dabei in die Begabungsforschung insofern integriert, als dass Expertentum oder Eminenz als Zielzustand bzw. finaler Abschnitt im Talententwicklungsprozess gesehen wird (s. Abbildung 1.1). Der statische Begabungsbegriff wird durch einen dynamischen Talentbegriff ersetzt. Leistungsstärke kann innerhalb dieser Modelle als Maß für Begabung in einem bestimmten Entwicklungsabschnitt und einer bestimmten fachlichen Domäne begriffen werden (Subotnik et al., 2011).

1.2.2. Weitere Perspektiven

Die kognitive Psychologie befasst sich mit den höheren geistigen Prozessen des Menschen, insbesondere mit der Art und Weise, wie Wissen erworben wird (Zimbardo & Gerrig, 2004). Aus kognitionspsychologischer Perspektive kann unter Schulleistung das in der Schule erworbene deklarative und prozedurale Wissen über Lerninhalte verstanden werden (Schrader & Helmke, 2008). Diese Lerninhalte werden in Bildungsstandards und Lehrplänen festgelegt. Durch

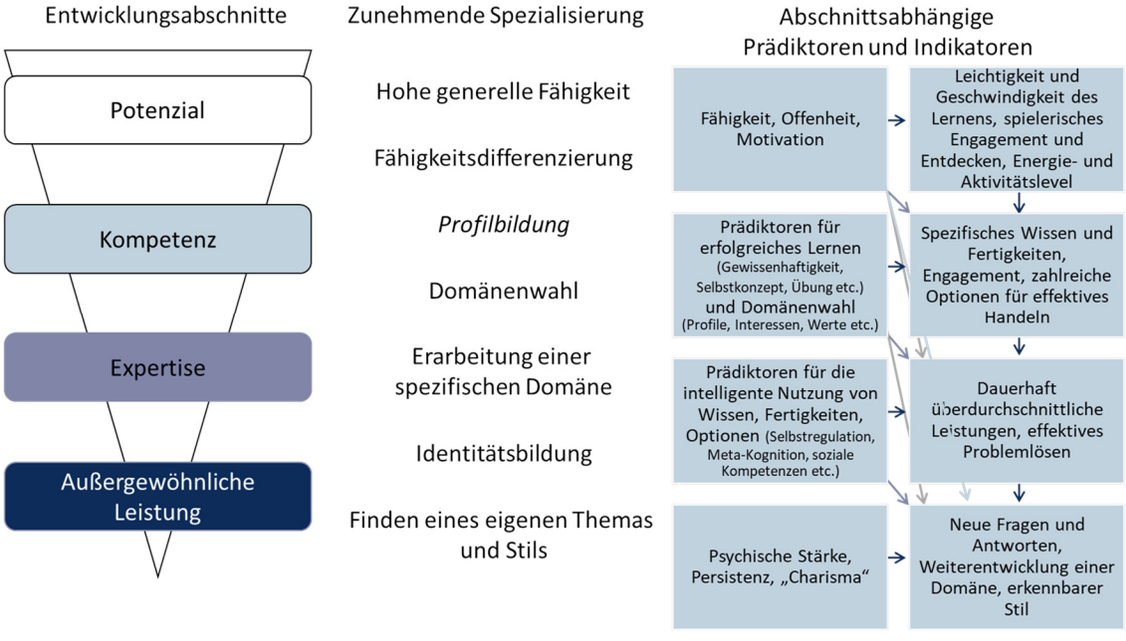


Abbildung 1.1. Das Talent Development in Achievement Domains Framework (TAD, Preckel et al., 2020).

curriculare Analyse können Wissensbereiche identifiziert werden, die der Leistungsfeststellung zugrunde liegen.

In Ergänzung zu dieser kognitionspsychologischen Sichtweise wird aus motivationspsychologischer Perspektive nicht nur das erworbene Wissen, also die Ergebnisse von Lernprozessen als Leistung bezeichnet, sondern auch die Lernprozesse selbst, sofern sie mit Anstrengung verbunden sind und anhand von anerkannten Gütemaßstäben beurteilt werden können (Ingenkamp & Lissmann, 2008; Klafki, 2007, Heckhausen, 1974). Dieser erweiterte Schulleistungsbegriff umfasst damit neben den kognitiven Leistungen auch sozio-emotionale und motivationale Komponenten (Lissmann & Jürgens, 2015).

Schließlich werden Schulleistungen und Kompetenzen häufig synonym verwendet. Kompetenzen definiert Weinert (2001, 27f) als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“. Dabei liegt die Betonung darauf, dass diese Kompetenzen bereichsspezifisch und weitgehend erlernbar sind, woraus einerseits ihre Abgrenzung zur Intelligenz als „weitgehend dekontextualisierte Denkleistung“ erwächst (Wilhelm & Nickolaus, 2013, S. 25). Andererseits geht der Kompetenzbegriff aber auch über den Wissensbegriff hinaus, da er als „Einheit von Wissen und Können gefasst“ wird (ebd.).

Ein gänzlich anderes Verständnis wird aus der Perspektive von Soziologie und Sozialpsychologie verfolgt, die das Individuum in seiner Wechselwirkung mit der sozialen Umwelt untersucht. Hier wird Leistung als soziales Konstrukt aufgefasst, welches in der Attribution durch Akteure des sozialen Systems entsteht (Budde, 2013; Kalthoff, 2000). Leistungsstark ist also, wer von den Mitmenschen als erfolgreich angesehen wird.

Es gibt schließlich Autoren, die in dem Begriff Schulleistung ein Sammelbecken für alle möglichen unterschiedlichen Konstrukte sehen, sodass die Definition des Begriffs am ehesten über seine Operationalisierung erfolgt (Langfeldt & Fingerhut, 1974).

1.2.3. Zusammenfassung

Im vorangegangenen Kapitel wurde zunächst herausgestellt, dass Leistung in den meisten betrachteten Bereichen auf spezielle Inhalte oder Fähigkeiten bezogen verstanden wird. Im Gegensatz zur Begabung, die als weitestgehend stabil und genetisch determiniert erachtet wird, wird Leistungsstärke als variabel, dynamisch und beeinflussbar durch multiple Faktoren auf Ebene

des Individuums und seines Umweltkontextes begriffen. Aus dieser multiplen Determiniertheit von Leistung, die sich beispielsweise auch in Rahmenmodellen der Schul- und Unterrichtsforschung widerspiegelt (Seidel, 2014), folgt, dass eine größere Vielzahl an Forschungsbereichen an der Erklärung von Leistungsexzellenz beteiligt sein sollte. Diese Vielzahl an Forschungsperspektiven führt potenziell zu einer größeren Vielfalt des Verständnisses und Herangehens an die Messung von Leistung und Leistungsstärke, was wiederum die Integration von Ergebnissen unterschiedlicher Forschungsarbeiten erschweren kann.

Literaturverzeichnis

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U. et al. (Hrsg.). (2002). *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Budde, J. (Hrsg.). (2013). *Unscharfe Einsätze: (Re-)Produktion von Heterogenität im schulischen Feld*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-19039-6

Ericsson, K. A. & Charness, N. (1994). Expert performance. Its structure and acquisition. *American Psychologist*, *49*(8), 725–747.

Ericsson, K. A., Roring, R. W. & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: an account based on the expert performance framework. *High Ability Studies*, *18*(1), 3–56. https://doi.org/10.1080/13598130701350593

Fend, H. (2008). *Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen* (2., durchgesehene Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gagné, F. (1985). Giftedness and Talent: Reexamining a Reexamination of Definitions. *Gifted Child Quarterly*, *29*(3), 103–112.

Hany, E. A. (2012). Zum Verhältnis von Begabung und Leistung. In A. Hackl, C. Pauly, O. Steenbuck & G. Weigand (Hrsg.), *Werte schulischer Begabtenförderung. Begabung und Leistung* (Karg-Hefte. Beiträge zur Begabtenförderung und Begabungsforschung, S. 35–40). Frankfurt, M.: Karg-Stiftung. https://doi.org/10.25656/01:9030

Harder, B. (2012). *Modelle zur Erklärung von Leistungsexzellenz im theoretischen und empirischen Vergleich* (Talentförderung, Expertiseentwicklung, Leistungsexzellenz, Bd. 13). Berlin: Lit.

Harder, B., Vialle, W. & Ziegler, A. (2014). Conceptions of giftedness and expertise put to the empirical test. *High Ability Studies*, *25*(2), 83–120. https://doi.org/10.1080/13598139.2014.968462

Heller, K. A. (2001). Teil I: Projektziele, Untersuchungsergebnisse und praktische Konsequenzen. In K. A. Heller (Hrsg.), *Hochbegabung im Kindes- und Jugendalter* (2. Aufl., S. 22–40). Göttingen: Hogrefe.

Kalthoff, H. (2000). „Wunderbar, richtig“. Zur Praxis mündlichen Bewertens im Unterricht. *Zeitschrift für Erziehungswissenschaft*, *3*(3), 429–446. https://doi.org/10.1007/s11618-000-0042-3

Kultusministerkonferenz. (2010). *Förderstrategie für leistungsschwächere Schülerinnen und Schüler. Beschluss vom 4.3.2010* [Support strategy for underachieving students] (Beschlüsse der Kultusministerkonferenz). Kronach, Oberfr: Carl Link.

Langfeldt, H. P. & Fingerhut, W. (1974). Empirische Ansätze zur Aufklärung des Kontruktes "Schulleistung". In K. A. Heller (Hrsg.), *Leistungsbeurteilung in der Schule.* Heidelberg: Quelle & Meyer.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life. First results from PISA 2000*. Paris: OECD Publishing. https://doi.org/10.1787/9789264195905-en

Perleth, C. (2001). Follow-up-Untersuchungen zur Münchner Hochbegabungsstudie. In K. A. Heller (Hrsg.), *Hochbegabung im Kindes- und Jugendalter* (2. Aufl., S. 358–446). Göttingen: Hogrefe.

Preckel, F., Golle, J., Grabner, R., Jarvin, L., Kozbelt, A., Müllensiefen, D. et al. (2020). Talent Development in Achievement Domains: A Psychological Framework for Within- and Cross-Domain Research. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, 691–722. https://doi.org/10.1177/1745691619895030

Schrader, F.-W. & Helmke, A. (2008). Determinanten der Schulleistung. In M. K.W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion* (S. 285–302). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91104-5_11

Schweizer, K. (2006). Intelligenz. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik. Mit 18 Tabellen* (S. 2–15). Berlin, Heidelberg: Springer Medizin Verlag Heidelberg.

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie. Integration von Struktur- und Prozessparadigma. https://doi.org/10.25656/01:14686

Subotnik, R. F., Olszewski-Kubilius, P. & Worrell, F. C. (2011). Rethinking Giftedness and Gifted Education: A Proposed Direction Forward Based on Psychological Science. *Psychological Science in the Public Interest : a Journal of the American Psychological Society*, *12*(1), 3–54. https://doi.org/10.1177/1529100611418056

Terman, L. M. (1925). *Genetic Studies of Genius. Mental and Physical traits of a thousand gifted children*. Stanford University Press.

Terman, L. M. (1954). The discovery and encouragement of exceptional talent. *American Psychologist*, *9*, 221–230.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Wilhelm, O. & Nickolaus, R. (2013). Was grenzt das Kompetenzkonzept von etablierten Kategorien wie Fähigkeit, Fertigkeit oder Intelligenz ab? *Zeitschrift für Erziehungswissenschaft*, *16*(S1), 23–26. https://doi.org/10.1007/s11618-013-0380-6

Zimbardo, P. G. & Gerrig, R. J. (2004). *Psychologie* (16., aktualisierte Aufl.). München: Pearson Studium.

2

Wie wird Leistung gemessen?

1. Wie wird Leistung gemessen?

Die konkrete Messung schulischer Leistung findet in verschiedenen gesellschaftlichen Kontexten mit unterschiedlichen Zielsetzungen statt. So sind Leistungsmessungen Grundlage von Bildungssteuerung auf politischer Ebene, sie liefern aber auch die Grundlage für das adaptive Angebot an Lerngelegenheiten und die individuelle Leistungsrückmeldung im Unterricht, sowie für die leistungsabhängige Verteilung von Bildungs- und beruflichen Chancen, beispielsweise indem leistungsstarke Schülerinnen und Schüler für bestimmte Förderprogramme ausgewählt werden oder durch die notengesteuerte Verteilung in unterschiedliche weiterführende Bildungsgänge. Im Folgenden wird genauer darauf eingegangen, wie die Messung von Leistungsstärke in diesen Bereichen erfolgt und welche Auswirkungen dies hat.

1.1. Leistungsstärke-Indikatoren für die Politik: PISA, Bildungstrend etc.

Das Ziel nationaler Regierungen ist es, ein leistungsfähiges Bildungssystem zu entwickeln. Dies soll auf der einen Seite dafür sorgen, dass alle Bürgerinnen und Bürger eine grundlegende Bildung erhalten, die es ihnen ermöglicht, sich an gesellschaftlichen Prozessen zu beteiligen. Darüber hinaus sind jedoch gut ausgebildete Menschen notwendig, um Deutschland als Wirtschaftsstandort wettbewerbsfähig zu halten und damit den Wohlstand zu sichern. Eine frühzeitige Rückmeldung darüber, wie gut beide Ziele erreicht werden, erhoffen sich die Kultusministerien über die Teilnahme an international vergleichenden Schulleistungsstudien (wie dem Programme of International Student Assessment (PISA), der Trends in International Mathematics and Science Study (TIMSS), der Internationalen Grundschul-Leseuntersuchung (IGLU/PIRLS) und der Internationalen Computer and Information Literacy Study (ICILS)) und durch das Bildungsmonitoring auf nationaler Ebene (IQB-Bildungstrends). Diese Studien finden in regelmäßigem Turnus statt und erlauben so die Beobachtung von Entwicklungstrends und die Evaluation von Bildungsinterventionen, wie zum Beispiel Schulstrukturreformen oder Investitionen in Bildungsprogramme. Abbildung 2.1 zeigt den Testturnus dieser nationalen und internationalen Assessments. In diesen repräsentativen Large-Scale-Assessment-Studien werden regelmäßig die Kompetenzen von Schülerinnen und Schülern in unterschiedlichen Fachdomänen gemessen.

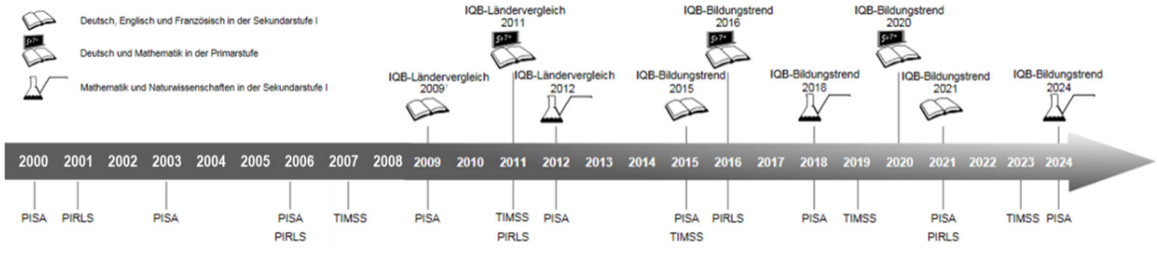


Abbildung 2.1. Zeitpunkte der Datenerhebungen der nationalen und internationalen Schulleistungsstudien von 2000 bis 2024. Basiert auf Stanat, Schipolowski & Pant (2019).

In Deutschland geben die Bildungsstandards vor, welche Kompetenzen Schülerinnen und Schüler in einem bestimmten Bildungsabschnitt erworben haben sollten. Dieser Kompetenzstand wird regelmäßig im Rahmen der Bildungstrends (ehemals Ländervergleiche) des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) gemessen. Die Fähigkeitsskala wird dabei in Kompetenzniveaus ansteigender Schwierigkeit unterteilt; Unterschiede zwischen Kompetenzstufen korrespondieren aber auch mit qualitativen Unterschieden in den zu ihrem Erreichen notwendigen Teilkompetenzen (s. z. B. Pant, Böhme, Stanat, Schipolowski & Köller, 2019). Damit können die Kompetenzstufen als kriteriale Normen angesehen werden. Die IQB-Bildungstrends definieren fünf Kompetenzstufen. Die höchste Kompetenzstufe, der sogenannte *Optimalstandard*, bezeichnet dabei Kompetenzen, die „bei sehr guten oder ausgezeichneten individuellen Lernvoraussetzungen und der Bereitstellung gelingender Lerngelegenheiten innerhalb und außerhalb der Schule erreicht werden können und die bei Weitem die Erwartungen der KMK-Bildungsstandards übertreffen“ (Pant et al., 2019).

Die Anteile von Schülerinnen und Schülern auf der höchsten Kompetenzstufe unterscheiden sich zwischen den verschiedenen Fächern, da die Einteilung in Kompetenzstufen fachspezifischen Kriterien folgt. Im letzten Bildungstrend der Grundschule lag der Anteil leistungsstarker Schülerinnen und Schüler, welche die oberste Kompetenzstufe erreichten, bei etwa 10 Prozent in Deutsch und 13 Prozent in Mathematik (Kohrt, Haag & Stanat, 2017; Weirich, Wittig & Stanat, 2017). In der Sekundarstufe I lagen die Anteile von Schülerinnen und Schülern, die den Optimalstandard erreichen, in Deutsch im einstelligen Bereich (Lesen: 3 %, Zuhören: 9 % und Orthografie: 7 %), in Englisch (Lesen und Zuhören) bei 8 bzw. 13 Prozent (Schipolowski et al., 2016), in Mathematik bei 4 Prozent und auch in den naturwissenschaftlichen Teilkompetenzen größtenteils im niedrigen einstelligen Bereich (Stanat, Schipolowski, Mahler, Weirich & Heschel, 2019). Alle Erhebungen liegen inzwischen im Trend mit zwei Messzeitpunkten vor, wobei sich der Anteil Leistungsstarker im Fach Englisch deutlich von 2009 bis 2015 vergrößert hat. In allen anderen Bereichen lag zwischen den zwei Messungen insgesamt ein rückläufiger Trend im Anteil leistungsstarker Schülerinnen und Schüler vor. Die Streuungen zwischen den Bundesländern waren dabei aber teils massiv, wobei sich die Anteile von Schülerinnen und Schülern, welche den Optimalstandard erreichten, zwischen den Ländern um bis zu 14 Prozentpunkte unterschieden.

Die Kompetenzstufen der internationalen Bildungsvergleichsstudien sind nicht direkt mit dem nationalen Kompetenzstufenmodell vergleichbar. Zum einen unterscheidet sich das Konzept der gemessenen Kompetenzen. Bei den internationalen Vergleichsstudien wird *Literacy* untersucht, welche einen stärker funktionalen Fokus setzt und zentrale grundlegende Kompetenzen untersucht, die eine Basis für gesellschaftliche Teilhabe und lebenslanges Lernen bilden (Weis &

Reiss, 2019) und nicht notwendigerweise mit dem deutschen Curriculum korrespondieren. Die Kompetenzstufen sollen dabei ebenfalls eine inhaltliche Interpretation erlauben, sind aber stärker unter statistischen Aspekten der Leistungsverteilung konstruiert worden als die Kompetenzstufen der IQB-Bildungstrends, welche curricularen, fachdidaktischen und lernpsychologischen Kriterien folgen (Pant et al., 2019; *PISA 2009 Technical Report,* 2012).

Die Anteile von Schülerinnen und Schülern auf den beiden höchsten Kompetenzstufen[1] (V und VI) bei PISA betrugen in den vergangenen Jahren 11 % im Lesen, 13 % in Mathematik und 10 % in den Naturwissenschaften. Diese Anteile waren signifikant höher als der OECD-Durchschnitt, aber einige OECD-Staaten hatten auch einen noch größeren Anteil von Schülerinnen und Schülern auf den höchsten Kompetenzstufen (die Maxima lagen im Lesen bei 15 %, in Mathematik bei 21% und in den Naturwissenschaften bei 13 %), während die nicht-OECD Staaten Singapur und China[2] jeweils die Spitze aller teilnehmenden Staaten bildeten - mit maximal 26, 44 bzw. 32 Prozent (Reinhold, Reiss, Diedrich, Hofer & Heinze, 2019; Schiepe-Tiska, Rönnebeck & Neumann, 2019; Weis et al., 2019). Bei computer- und informationsbezogenen Kompetenzen, die in ICILS gemessen werden, befinden sich lediglich knapp zwei Prozent der deutschen Schülerinnen und Schüler auf dem obersten Kompetenzniveau, während in Korea etwa neun Prozent Kompetenzstufe V erreichen (Eickelmann, Bos, Gerick & Labusch, 2019).

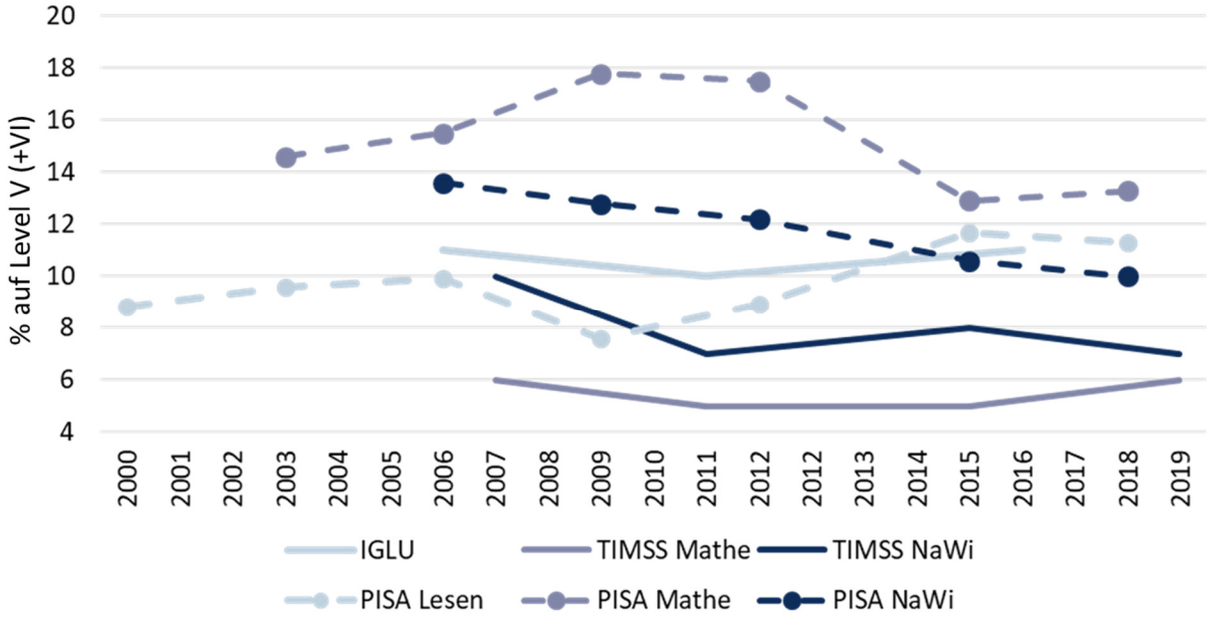


Abbildung 2.2. Anteil von Schülerinnen und Schülern, die die oberste(n) Kompetenzstufe(n) in den internationalen Bildungsvergleichsstudien erreichten. Daten sind den Berichten entnommen.

[1] Zu Beginn wurde die Kompetenzskala in lediglich 5 Kompetenzniveaus unterteilt. Ab dem Jahr 2009 wurde die Skala um eine Stufe nach oben und eine Stufe nach unten erweitert, um Unterschiede an den Enden der Leistungsverteilung noch differenzierter abbilden zu können.

[2] Zusammenschluss der Verwaltungseinheiten Peking, Shanghai, Zhagsu und Zhejiang

Bei den Grundschulstudien IGLU und TIMSS lag der Anteil an Kindern auf der obersten Kompetenzstufe bei 11 Prozent im Lesen und 6 bzw. 7 Prozent in Mathematik bzw. Naturwissenschaften (s. Abbildung 2.2). Die Werte in Lesen und den Naturwissenschaften entsprachen in etwa dem europäischen Durchschnitt. In Mathematik lag der Anteil Leistungsstarker deutlich unter dem Mittelwert der EU-Staaten (9 %). Verglichen mit den Staaten an der Spitze (sowohl international als auch bezogen auf die EU) sind die Anteile leistungsstarker Schülerinnen und Schüler in Deutschland viel geringer (in Lesen erreichen z.B. 26 Prozent in Russland und 21 Prozent der Schülerinnen und Schüler in Irland die oberste Kompetenzstufe; in Mathematik sind es 37 % in Taiwan und 15 % in Irland und in den Naturwissenschaften 29% in Korea und 15% in Finnland und Bulgarien).

In den internationalen Untersuchungen wird zuweilen auch die die erreichte Punktzahl am 95. Perzentil als Maß für die Effektivität der Spitzenförderung eines Landes verglichen. Schaut man sich für diesen Indikator den längerfristigen Trend an, so ist zu konstatieren, dass die Kompetenzen deutscher Grundschülerinnen und -schülern am oberen Ende der Leistungsverteilung im Lesen und in Mathematik zuletzt zugenommen haben, während in den naturwissenschaftlichen Kompetenzen und in anderen Altersgruppen ein eher negativer oder stagnierender Trend zu beobachten ist (s. Abbildung 2.3).

Auf die Ergebnisse dieses Bildungsmonitorings berief sich die Kultusministerkonferenz im Jahr 2015, als sie die „Strategie zur Förderung leistungsstarker und potenziell besonders leistungsstarker Schülerinnen und Schüler“ beschloss. Darin stellte sie fest:

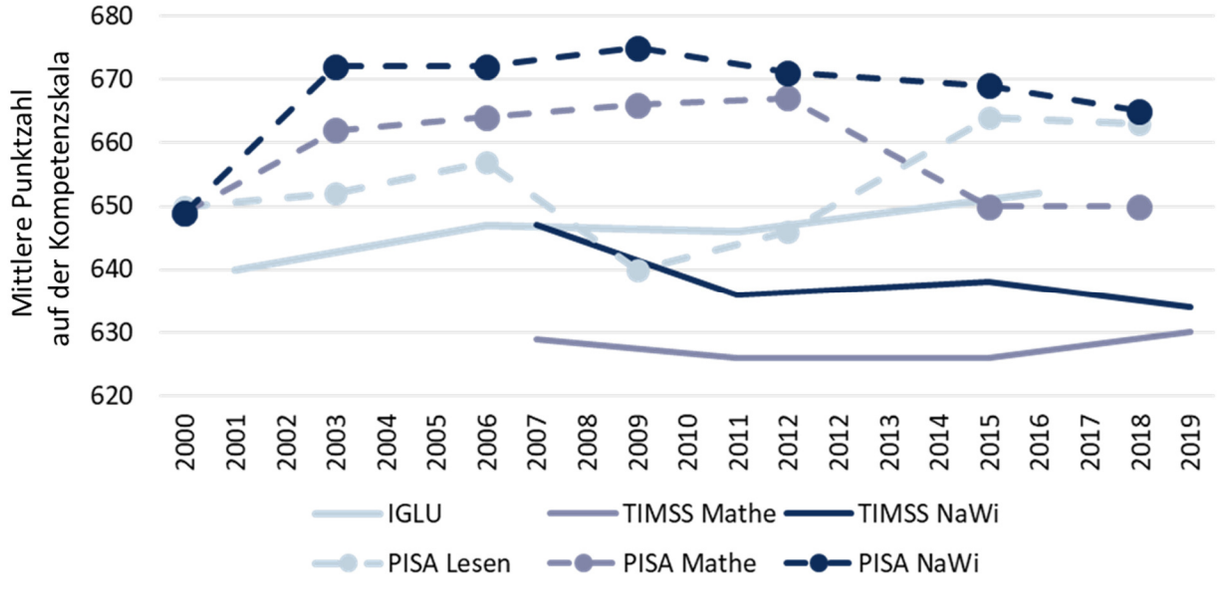


Abbildung 2.3. Testleistung von Schülerinnen und Schülern am 95. Perzentil in den internationalen Bildungsvergleichsstudien. Daten sind den Berichten entnommen.

„Ein Blick auf den vergleichsweise geringen Anteil von Schülerinnen und Schülern auf den beiden oberen Kompetenzstufen der PISA-Studien bzw. der Ländervergleiche der Kultusministerkonferenz sowohl im Bereich der Naturwissenschaften/Mathematik als auch in Deutsch und Englisch verdeutlicht die Notwendigkeit, die Förderung von leistungsstarken und potenziell leistungsfähigen Schülerinnen und Schülern zu verbessern.“ (Kultusministerkonferenz, 2015, S. 3)

Ein besonderer Fokus auf leistungsstarke Schülerinnen und Schüler findet sich daraufhin in den nationalen Bildungsvergleichsstudien im Jahr 2016, für das der Berichtsband ein eigenes Kapitel zur Leistungsspitze umfasste (Neuendorf, Kuhl & Jansen, 2017).

Weiterhin wurde in der Folge durch den Bund und die Länder eine Initiative gestartet, welche diese Anteile erhöhen und Schülerinnen und Schüler mit hohen Leistungen bzw. guten Lernvoraussetzungen intensiver fördern sollte (Bundesministerium für Bildung und Forschung [BMBF] & Kultusministerkonferenz [KMK], 2016). Diese Initiative stellte Mittel bereit, durch die in 300 Schulen deutschlandweit Konzepte, Maßnahmen und Materialien unter wissenschaftlicher Begleitung entwickelt werden und in einer zweiten Phase in die Breite der gesamten Schullandschaft getragen werden sollen. Das damit begründete Projekt LemaS („Leistung macht Schule“) startete im Jahr 2018. Dieses Projekt stand zu Beginn vor der Herausforderung, Begriffe von Leistungsstärke und Leistungspotenzial zu entwickeln, welche Zustimmung sowohl in der schulischen Praxis als auch in der Politik finden, Diskurse und Forschungsansätze unterschiedlicher Wissenschaftsdisziplinen integrieren und eine Grundlage für die wissenschaftliche Begleitung und Evaluation darstellen konnte. Das Verständnis von Begabung und Leistung wurde in diesem Zusammenhang sehr weit gehalten, um einen Konsens möglich zu machen. Der resultierende Leistungsbegriff wird als mehrdimensional und entwicklungsbezogen beschrieben, der einerseits schulbezogene Leistung, aber andererseits auch Persönlichkeitsentwicklung, Lebenskontext und gesellschaftlich verantwortliches Handeln einbezieht (Forschungsverbund LemaS, 2018; Weigand, 2020).

1.2. Leistungsdiagnostik in der Praxis

In der schulischen Praxis findet die Leistungsdiagnostik im Unterricht in Form von Notengebung durch die Lehrkräfte statt. Mittels Schulnoten überführen Lehrkräfte die zu bewertenden Leistungs- und Verhaltensaspekte in Zahlen (Maaz, Baeriswyl & Trautwein, 2013). Zwar gibt es unterschiedliche Ansichten dazu, welche Verhaltensaspekte in welchem Maße in die Bewertung von Schulleistung Eingang finden sollten. Auch wird an Noten kritisiert, dass sie stark von Referenzgruppeneffekten betroffen sind (Ingenkamp, 1977; Rheinberg, 2008) und verschiedene Lehrkräfte bei der Benotung ein- und derselben Arbeit oft mehrere Notenpunkte auseinanderliegen

(Lintorf, 2012). Dennoch gibt der Fakt, dass Lehrkräfte sich bezüglich der Rangreihe der Leistungen innerhalb einer Klasse in der Regel einig sind (Maaz, Baeriswyl & Trautwein, 2013), einen Hinweis darauf, dass es tatsächlich ein dahinterliegendes Leistungskonstrukt gibt, welches von Bewertern, also Lehrkräften, geteilt wird. Ein Dilemma besteht darin, dass Schulnoten unterschiedliche Funktionen zukommen (Lüders, 2001). So gibt es die gesellschaftliche Funktion, die beispielsweise darin besteht, beim Zugang zur nächsten Bildungsetappe Zuweisungen nach dem meritokratischen Prinzip vornehmen zu können und es gibt die pädagogische Funktion, die darin besteht, Schülerinnen und Schülern in motivierender und disziplinierender Weise Rückmeldungen über ihren Lernstand und ihr Arbeitsverhalten zu geben (Tent, 2006). Man spricht diesbezüglich auch von summativen versus formativen Bewertungen. Während die erste Funktion erfordert, dass Noten ausschließlich nach dem Leistungskriterium vergeben werden sollten, um Leistungsunterschiede zwischen Lernenden darstellen zu können, erfordert die zweite Funktion eine Orientierung an individuellen Voraussetzungen und Potenzialen. Mit den immer stärker standardisierten und zentralisierten Abschlussprüfungen verschiebt sich der Fokus von Leistungsbewertungen stärker in Richtung objektivierbarer, manifester Leistung. Damit folgt das deutsche Bildungssystem einem Trend, der in den meisten anderen Ländern bereits die Norm darstellt (Eurydice-Netz, 2009; Graf, Harych, Wendt, Emmrich & Brunner, 2016).

1.3. Identifikation von Leistungsstarken durch die Forschung

In der Forschung unterliegt die Identifikation von Leistungsstarken nach Ziegler (2018) vier Problemen – dem Anwendbarkeitsproblem, dem Indikatorproblem, dem Referenzproblem und dem Signifikanzproblem. Zunächst stellt sich die Frage, auf welchen Leistungsbereich man das Kriterium anwenden kann – welches sind also *Bereiche*, in denen man Leistungen sinnvoll messen kann? Hier wird einerseits häufig darauf verwiesen, dass es sich um eine Domäne handeln sollte, welcher gesellschaftlich ein Wert zugeschrieben werden kann. Damit wird die normative Ebene von Leistungsbewertungen betont. Daneben geht es aber auch um die Frage, ob überhaupt ein Gütemaßstab existiert, auf dem eine Leistung gemessen werden kann. Das *Indikatorproblem* bezieht sich auf die Frage, mit welchem Messinstrument Leistung in einer Domäne gemessen werden kann, welches also Kennzeichen von Leistung in einer Domäne sind. Im nächsten Schritt geht es um die Frage, welche *Bezugsnorm* bei der Messung und Beurteilung von Leistung angelegt werden sollte und beim *Signifikanzproblem* darum, wann eine Leistung als herausragend gilt (Ziegler, 2018, S. 1281). Jeder dieser Aspekte muss bei der Definition von Leistungsstärke in der Forschung adressiert werden. In welcher Form dies bisher geschieht, ist Gegenstand von Artikel I, der einen systematischen Überblick über die Definition von Leistungsstärke gibt.

1.4. Ableitung der Forschungsfragen von Artikel I und methodisches Vorgehen

Wie aus den obigen Darstellungen deutlich geworden ist, gibt es in einigen praktisch bedeutsamen Bereichen wie der Bildungspolitik und der Unterrichtspraxis bereits recht genaue Vorstellungen davon, wie Leistungsstärke gemessen werden sollte: im ersten Fall ist es die Verortung auf einer spezifischen Kompetenzstufe der Bildungsmonitoringstudien, im zweiten Fall ist es die Vergabe der Note 1, wenn ein Erwartungshorizont erreicht oder übertroffen wurde. Worauf beruft sich aber die wissenschaftliche Forschung, wenn sie sich mit exzellenten Leistungen von Schülerinnen und Schülern befasst? Welche Domänen, Indikatoren, Referenznormen und Signifikanzkriterien werden genutzt, um Leistungsstarke in der Forschung zu identifizieren und zu beschreiben? Dies wird in Artikel I dieser Dissertation näher beleuchtet. Die konkreten Forschungsfragen lauten: 1) Welche Indikatoren werden zur Operationalisierung von Leistungsstärke herangezogen? und 2) Welche Cut-off-Werte werden zur Bestimmung von Leistungsstärke festgelegt? Dabei bezieht sich die erste Fragestellung auf das Anwendbarkeits- und Indikatorproblem, während die zweite Fragestellung den Umgang von Forschenden mit dem Referenz- und dem Signifikanzproblem unter die Lupe nimmt. Hierzu wurde eine systematische Überblicksarbeit erstellt. In unterschiedlichen Fachdatenbanken wurde nach Artikeln in wissenschaftlichen Journals aus den Jahren 2000 bis 2020 gesucht, welche sich mit akademisch leistungsstarken Schülerinnen und Schülern befassten. Diese wurden nach den oben benannten Kriterien ausgewertet. Die Datengrundlage und der Analysecode zur Erstellung der Tabellen und Abbildungen finden sich unter https://osf.io/jzkv6/.

I

Wer ist leistungsstark? Operationalisierung von Leistungsstärke in der empirischen Bildungsforschung seit dem Jahr 2000

Claudia Neuendorf

Malte Jansen

Poldi Kuhl

Miriam Vock

Neuendorf, C., Jansen, M., Kuhl, P. & Vock, M. (2022). Wer ist leistungsstark? Operationalisierung von Leistungsstärke in der empirischen Bildungsforschung seit dem Jahr 2000. Zeitschrift für Pädagogische Psychologie.
https://doi.org/10.1024/1010-0652/a000343

Abstract

Leistungsstarke Kinder und Jugendliche sind in den letzten Jahren zunehmend in den Fokus der Bildungspolitik und der Bildungsforschung gerückt. Allerdings gibt es in der Forschung bislang kein geteiltes Verständnis darüber, was genau unter akademischer Leistungsstärke zu verstehen ist. Die vorliegende Arbeit gibt einen systematischen Überblick darüber, wie Forschende, die seit dem Jahr 2000 die Gruppe der leistungsstarken Schülerinnen und Schüler erforschten, Leistungsstärke in ihren Studien operationalisiert haben. Dabei wurde insbesondere untersucht, welche Leistungsindikatoren genutzt wurden, ob ein spezifischer Fachbezug hergestellt wurde und welche Cut-off-Werte und Vergleichsmaßstäbe angelegt wurden. Die systematische Datenbanksuche lieferte insgesamt $N = 309$ Artikel, von denen $n = 55$ die Einschlusskriterien erfüllten. Die Ergebnisse zeigen, dass eine große Vielfalt in der Operationalisierung von Leistungsstärke vorliegt. Die meistgenutzten Leistungsindikatoren waren Noten und Testwerte, wobei fächerübergreifende und fachspezifische Definitionen beide häufig waren. Die Cut-off-Werte der Studien waren zum Teil schwierig vergleichbar, aber dort, wo ein Populationsbezug hergestellt werden konnte, lag der Median des Populationsanteils Leistungsstarker bei 10 Prozent. Die Studie diskutiert methodische und inhaltliche Rahmenbedingungen, welche sich auf die Operationalisierung von Leistungsstärke und ihre Vergleichbarkeit über Studien hinweg auswirken. Die vorliegende Arbeit schließt mit Empfehlungen zur Operationalisierung von Leistungsstärke.

Schlüsselbegriffe: Leistungsstarke Schüler*innen, Operationalisierung, Definition, Review, Hochbegabung

Einleitung

Leistungsstarke Schülerinnen und Schüler sind in den letzten Jahren zunehmend in den Fokus von Bildungspolitik, Bildungsforschung und Bildungspraxis geraten. So wurde im Jahr 2015 die Strategie zur Förderung leistungsstarker und potenziell leistungsstarker Schülerinnen und Schüler von der Kultusministerkonferenz [KMK] verabschiedet. Bund und Länder haben im Jahr 2016 zudem die gemeinsame Initiative zur Förderung leistungsstarker und potenziell besonders leistungsfähiger Schülerinnen und Schüler beschlossen. Ziel dieser Initiative, die derzeit unter dem Namen „LemaS“ („Leistung macht Schule“) läuft, ist es, in den Schulen Ansätze zu entwickeln, um leistungsstarke und potenziell leistungsstarke Kinder und Jugendliche künftig angemessener zu fördern und so auch ihren Anteil an der Schülerschaft zu erhöhen. Diese Entwicklung soll durch Begleitforschung unterstützt werden (Bundesministerium für Bildung und Forschung [BMBF] & KMK, 2016). Forschung, die sich mit der Gruppe der leistungsstarken Schülerinnen und Schüler beschäftigen will, unterliegt dabei der Notwendigkeit, sowohl einen theoretischen Begriff von Leistung zugrunde zu legen als auch eine Operationalisierung von Leistungsstärke zu finden. Um feststellen zu können, ob entwickelte Fördermaßnahmen ihre Ziele erreichen, muss somit eine begründete und nachvollziehbare Festlegung dazu getroffen werden, welche Kinder und Jugendlichen als leistungsstark kategorisiert werden. Auch bei der Rezeption wissenschaftlicher Literatur ist es wichtig, die jeweils zugrunde gelegte Definition von Leistungsstärke zu kennen, um einordnen zu können, ob sich die berichteten Ergebnisse zwischen Studien sinnvoll aufeinander beziehen lassen. Während sowohl hochbegabte als auch leistungsschwache Schülerinnen und Schüler schon viel untersucht wurden, ist die Auseinandersetzung mit leistungsstarken Schülerinnen und Schülern ein vergleichsweise junges, sich gerade erst entwickelndes Forschungsfeld (Köller & Baumert, 2017) und es finden sich bislang noch keine Konventionen zu ihrer Definition innerhalb der Wissenschaftscommunity.

Im vorliegenden Beitrag wird daher ein systematischer Überblick über deutschsprachige und internationale empirische Studien gegeben, die seit dem Jahr 2000 die Gruppe der leistungsstarken Schülerinnen und Schüler untersucht haben. Im Mittelpunkt der Übersichtsarbeit steht die Frage, welche Operationalisierung von Leistungsstärke Forschende in ihren Untersuchungen gewählt haben.

Das Verhältnis von Leistungsstärke und Begabung

Obwohl ein intuitives Verständnis dafür existiert, wann jemand exzellente Leistungen erbringt, liegt keine einheitliche wissenschaftliche Definition dazu vor, was unter Schulleistung zu verstehen ist (Brühwiler & Helmke, 2018). Forschungsarbeiten, die sich mit Kindern und Jugendlichen beschäftigen, welche herausragende Leistungen im akademischen Bereich zeigen,

beziehen sich häufig auf die Tradition der Begabungsforschung. Diese entstand zeitgleich mit der Entwicklung der ersten Intelligenztests und bezog sich entsprechend auf die Untersuchung und Beschreibung von Personen, die Höchstleistungen in kognitiven Grundfähigkeiten erbringen, wie sie typischerweise von Intelligenztests gemessen werden (Lubinski, 2016; Terman, 1954). In der Begabungsforschung war lange Zeit die Setzung weithin akzeptiert, die besten zwei Prozent einer Population bzw. Personen, die wenigstens zwei Standardabweichungen über dem Populationsmittelwert liegen, als hochbegabt zu definieren (Amelang & Schmidt-Atzert, 2006; Rost & Buch, 2018). Kognitive Begabung wird dabei als zu einem größeren Anteil genetisch determiniert angesehen (Galton, 1892; Terman, 1926, 1954). Konzeptionell davon abgegrenzt wurde häufig Leistungsstärke (also das Erbringen sehr guter Leistungen in der Schule bzw. in curriculumsnahen, bereichsspezifischen Tests), die sich aus hoher kognitiver Begabung ergeben kann, es aber nicht zwangsläufig tun muss (Hany, 2012). Die Gegenüberstellung von begabten und leistungsstarken Schülerinnen und Schülern erfolgte beispielsweise prominent im Marburger Hochbegabtenprojekt (Rost, 2009), in dem hochbegabte und hochleistende Schülerinnen und Schüler in ihrer Entwicklung miteinander verglichen wurden. In dieser Studie wurden aus der Hochleistungsstichprobe 12 Prozent der Jugendlichen ausgeschlossen, da sie gleichzeitig einen IQ von über 130 hatten, also das Hochbegabungskriterium erfüllten.

Aktuelle Begabungsmodelle, wie das Megamodell von Subotnik, Olszewski-Kubilius und Worrell (2011), schlagen ein verändertes Konzept von Hochbegabung vor, indem sie von der einseitigen Fokussierung auf Intelligenz Abstand nehmen. Stattdessen wird der Prozess der Talententwicklung beschrieben, welcher sich in einer Entwicklung von allgemeineren kognitiven Grundfähigkeiten und Prädispositionen hin zu immer spezialisierterem Wissen und Können vollzieht, um schlussendlich Expertise in einer Domäne hervorzubringen. Leistungsstärke ist aus dieser Perspektive sowohl Resultat früherer Investitionen in ein Talent als auch Prädiktor für eine erfolgreiche Weiterentwicklung des Talents bei entsprechender Förderung. Leistungsstärke wird damit als Maß für Begabung in einem bestimmten Entwicklungsabschnitt und einer bestimmten fachlichen Domäne begriffen. Für die Erforschung von Talententwicklung, wie sie beispielsweise im TAD-Framework (Talent Development in Achievement Domains Framework, Preckel et al., 2020) angeregt wird, ist die Bestimmung von fachbezogener Leistungsstärke damit ähnlich zentral wie die Messung breit angelegter kognitiver Begabung. Um Ergebnisse von Talententwicklungsprozessen zu bewerten und Vorhersagen über den weiteren Erfolg in einer bestimmten Talentdomäne zu machen, stellt sich daher die Frage, wie Leistungsstärke in einem Fach operationalisiert werden soll.

Diese neuen Entwicklungen in der Begabungsforschung anerkennend, wird aus forschungspraktischen Gründen in der vorliegenden Arbeit dennoch eine Abgrenzung zwischen Leistungs- und Begabungskonzept getroffen, wie sie beispielsweise in den Moderatorenmodellen von Gagné (1985) oder Heller (2001) angelegt ist. Dabei wird Begabung als weitgehend angeborenes, bereits in jungem Alter vorhandenes, hohes allgemeines kognitives Fähigkeitspotenzial verstanden. Leistungsstärke hingegen wird in seinem engen Verständnis verwendet und auf bereits realisierte hohe Leistungen im akademischen Bereich beschränkt.

Empirische Vorarbeiten

Auch wenn es sich bei kognitiven Fähigkeitstests um prinzipiell kontinuierliche Maße handelt, hatten sich historisch in der Begabungsforschung Kriterien für die kategoriale Definition von Hochbegabung (2%, s.o.) entwickelt. Ebenso existieren Kriterien für die Diagnostik von Teilleistungsschwächen und sonderpädagogischen Förderschwerpunkten. Für fachbezogene schulische Leistungsstärke hingegen liegen keine Konventionen vor und bisher existiert noch keine Forschungsarbeit, die eine Übersicht darüber bietet, wie Forschende Leistungsstärke definieren und operationalisieren – d.h. welche Kategorisierungsstrategien sie nutzen, um leistungsstarke Schülerinnen und Schüler in ihrer Forschung zu identifizieren. Ein Beispiel für die Diversität an Definitionen von Leistungsstärke bieten nationale und internationale Large-Scale-Assessments, in denen Aussagen über die Leistungsfähigkeit von Schülerinnen und Schülern beziehungsweise deren Anteile an der Schülerschaft getroffen werden. Trotz der hohen Standardisierung legen auch diese Studien bei ihrer Interpretation der Leistungen teilweise unterschiedliche Operationalisierungen von Spitzenleistungen zugrunde. Zum Teil ist dies ein Resultat unterschiedlicher Kompetenzstufenmodelle bei den unterschiedlichen Assessments: So ist die Kompetenzskala beim IQB-Bildungstrend in fünf inhaltlich definierte Kompetenzbänder unterteilt, während die Kompetenzstufen bei PISA im Lesen von ursprünglich fünf Stufen auf sieben Stufen erweitert wurden, um im oberen und unteren Bereich besser zu differenzieren. Die ursprüngliche fünfte Kompetenzstufe wurde dabei nochmals unterteilt. Diese Anpassung verdeutlicht bereits, dass eine Verschiebung des Interesses an leistungsstarken Schülerinnen und Schülern stattgefunden hat. In der überwiegenden Zahl der Berichte werden zum einen oberen fünf Prozent und zum anderen der Anteil der Schülerinnen und Schüler auf Kompetenzstufe V (bzw. V und VI bei PISA) als leistungsstark, hochkompetent oder „Spitzengruppe“ bezeichnet (z. B. OECD, 2009). Mitunter gibt es allerdings auch Abweichungen von diesem Vorgehen. In einem Unterkapitel des nationalen Berichts zu PISA 2006 werden die 25 Prozent leistungsstärksten in den Naturwissenschaften „hochkompetent“ genannt (Prenzel, Schütte & Walter, 2007). In Zusatzanalysen zu PISA 2012 werden die besten zehn Prozent in Mathematik bzw.

Naturwissenschaften als „Highest-achieving students“ herausgestellt (OECD, 2015). Bei IGLU 2011 (Bos et al., 2012) wurden fachübergreifende Fähigkeitsprofile errechnet und die beiden obersten Profile unter der Bezeichnung „Schülerinnen und Schüler mit hohen Leistungen“ beschrieben.

Systematische Untersuchungen zur Operationalisierung hoher Leistungen bezogen sich bisher meist auf das traditionelle Begabungskonstrukt auf Grundlage kognitiver Fähigkeiten. In dieser Begabungsforschung existieren beispielsweise Arbeiten, die sich mit Identifikationsverfahren für Begabtenförderprogramme beschäftigten (Acar, Sen & Cayirdag, 2016; McBee, 2006; Rothenbusch, Zettler, Voss, Lösch & Trautwein, 2016). Zudem liegen zwei Übersichtsarbeiten vor, die die Operationalisierung von Begabung in Forschungsarbeiten systematisch untersucht haben (Carman, 2013; Ziegler & Raul, 2000). In diesen Übersichtsarbeiten wurde deutlich, dass Intelligenztests und Leistungstests beziehungsweise Schulleistung die häufigsten Methoden der Identifikation Begabter für wissenschaftliche Studien waren, gefolgt von Nominierungen, zum Beispiel durch die Lehrkräfte. Häufig wurden in den Studien mehrere Indikatoren herangezogen, um Begabung festzustellen. Bemerkenswert ist aber auch der Befund, dass sich ein substanzieller Anteil an Studien fand, die keine oder nur ungenügende Angaben zur Operationalisierung von Begabung machten.

Herleitung der Forschungsfragen

Zur Beantwortung der Fragestellung, wie Leistungsstarke in Studien identifiziert wurden, wird im vorliegenden Beitrag eine systematische Übersichtsarbeit erstellt. Ziel war es, alle Artikel seit dem Jahr 2000, welche leistungsstarke Schülerinnen und Schüler untersuchten, zu identifizieren und deren Operationalisierung des Konstrukts Leistungsstärke zu systematisieren.

Den zwei im Folgenden dargestellten Teilfragestellungen wird hierbei nachgegangen:

Forschungsfrage 1: Welche Indikatoren werden zur Operationalisierung von Leistungsstärke herangezogen?

Die erste Teilfrage, die in dieser Arbeit beantwortet werden soll, ist, welche Art von Leistungsmaß zur Identifikation leistungsstarker Schülerinnen und Schüler eingesetzt wird (Frage 1.a). Das in der Praxis bedeutsamste Schulleistungsmaß stellen Schulnoten dar. Mittels Schulnoten überführen Lehrkräfte die zu bewertenden Leistungs- und Verhaltensaspekte in Zahlen (Maaz, Baeriswyl & Trautwein, 2013). Sie erfüllen im Bildungssystem verschiedene Funktionen (z. B. pädagogische Funktion, Selektions- und Allokationsfunktion), weshalb es nicht verwunderlich ist, dass neben der schulischen Leistung auch andere Aspekte wie das Arbeitsverhalten und die Motivation in Noten einfließen und nicht nur kriteriale, sondern auch individuelle und soziale Referenzmaßstäbe

bei der Vergabe eine Rolle spielen (Hochweber, 2010; Lintorf, 2012a, 2012b; Rüdiger, Jansen & Rjosk, 2021). Daher ist die Vergleichbarkeit von Noten über einzelne Klassen, Schulen und Schulformen hinweg nur bedingt gegeben. Den Schulnoten werden häufig objektive und standardisierte Leistungs- und Kompetenztests gegenübergestellt. Hierzu gehören zum Beispiel Kompetenztests aus nationalen und internationalen Large-Scale-Assessments, die dem Monitoring und der Evaluation des Bildungssystems dienen (z. B. Stanat, Schipolowski, Mahler, Weirich & Henschel, 2019; Reiss, Weis & Klieme, 2019; Schwippert et al., 2020), aber auch weitere normierte und standardisierte Tests, die sich auch zur Individualdiagnostik eignen (z.B. ELFE II, Lenhard, Lenhard & Schneider, 2020, DEMAT 2+, Krajewski, Dix & Schneider, 2020, BASIS-MATH, Moser Opitz, Stöckli, Grob, Nührenbörger & Reusser, 2019). Es ist anzunehmen, dass die Operationalisierung von Leistungsstärke durch Noten und Kompetenz- und Leistungstests einen großen Anteil der Studien charakterisieren wird. Bei den Überblicksarbeiten aus dem Bereich der Begabungsforschung waren Nominierungen durch Lehrkräfte ebenfalls zu einem kleineren Anteil vertreten.

Ob häufiger Einzelindikatoren oder häufiger – wie bei Ziegler und Raul (2000) – eine Kombination mehrerer Indikatortypen (z.B. Noten und Testleistungen) eingesetzt werden, ist eine weitere Frage, die untersucht werden soll (Frage 1.b). McBee und Makel (2019) wiesen darauf hin, dass die Kombination mehrerer Indikatoren zu einem Hochbegabungs-Index, je nach Korrelation untereinander und genutzter Kombinationsregel, einen bedeutsamen Einfluss auf den Anteil Hochbegabter hat. Gleichzeitig wird häufig empfohlen, in der Diagnostik multiple Instrumente einzusetzen, um eine höhere Reliabilität zu erreichen (Amelang & Schmidt-Atzert, 2006). Gleiches gilt für die Operationalisierung von Leistungsstärke.

Weitergehend soll die Domänenspezifität des Konstrukts Leistungsstärke untersucht werden. Im Gegensatz zur traditionellen Begabungsforschung, in der Intelligenz als fachunabhängige kognitive Leistungsfähigkeit im Fokus steht, muss Leistung im Sinne von Performanz sich erst in einem oder mehreren Fächern manifestieren, um festgestellt zu werden. Dabei ist es nicht ungewöhnlich, dass Leistungen einer Schülerin oder eines Schülers in verschiedenen Fächern unterschiedlich ausfallen. So erreichte beispielsweise im Bildungstrend 2016 etwa ein Drittel der leistungsstarken Grundschülerinnen und Grundschüler (30 %, bzw. 7 % der Gesamtpopulation) sowohl in Mathematik als auch in Deutsch die höchste Kompetenzstufe in beiden Fächern. 27 Prozent waren hingegen nur in Mathematik und 43 Prozent nur im Fach Deutsch leistungsstark (Neuendorf, Kuhl & Jansen, 2017). Eine weitere Studie, welche leistungsstarke Jugendliche anhand von PISA-Daten der Jahre 2000 und 2003 untersuchte, stellte fest, dass lediglich etwa drei Prozent der Population in allen drei Inhaltsdomänen (Lesen, Mathematik und Naturwissenschaft)

zum leistungsstärksten Zehntel gehörten (Zimmer, Brunner, Lüdtke, Prenzel & Baumert, 2007). Der fachliche Bezug wird auch im TAD-Framework (Preckel et al., 2020) hervorgehoben, welches die Erforschung und Förderung von Leistungsentwicklung in spezifischen Leistungsdomänen zum Ziel hat.

Gleichzeitig kann Hochleistung auch als Gegenstück zur Hochbegabung verstanden und damit ebenfalls domänenübergreifend konzeptualisiert werden (Rost, 2009). Zum Beispiel untersuchten Zimmer et al. (2007) „vielseitig hochkompetente“ Schülerinnen und Schüler, die in allen Kompetenzbereichen (Mathematik, Lesen und Naturwissenschaften) zu den Leistungsstärksten gehörten. Mit Verweis auf die hohen Korrelationen zwischen Leistungen in verschiedenen Fächern definierten Köller und Baumert (2017) für die BERLIN-Studie Leistungsstärke aufgrund eines fächerübergreifenden Faktorscores. Damit stellt sich für die vorliegende Übersichtsarbeit die Frage, ob Forschende Leistungsstärke eher fachspezifisch oder fächerübergreifend definieren (Frage 1.c).

Forschungsfrage 2: Welche Cut-off-Werte werden zur Bestimmung von Leistungsstärke herangezogen?

Carman (2013) stellte in seiner Untersuchung fest, dass ein Großteil der Studien im Bereich der Begabungsforschung – obwohl für die kognitive Hochbegabung mit dem 2%-Kriterium zumindest eine teilweise geteilte Konvention vorlag - entweder keine Cut-off-Werte oder keine Verteilung der Testwerte in einer Normstichprobe angegeben hatte und auch das genaue Testinstrument selten berichtet wurde. Da damit häufig kaum nachzuvollziehen ist, wer als begabt kategorisiert wurde, sei eine Interpretierbarkeit der Ergebnisse und deren Übertragbarkeit auf andere Stichproben gefährdet. Diesem zentralen Kritikpunkt soll auch in der vorliegenden Untersuchung in Bezug auf Leistungsstärke nachgegangen werden. Dazu werden die Studien dahingehend kodiert, ob und welche Cut-off-Werte für die jeweiligen Indikatoren berichtet werden. In diesem Zusammenhang besonders relevant ist weiterhin die Frage, welche Art von Bezugsnorm verwendet wird. Bei einer kriterialen Bezugsnorm etwa würden absolute Notenstufen, Testwerte oder Kompetenzstufen verwendet. Bei einer sozialen Referenznorm wird die relative Position in der Leistungsverteilung genutzt. Dazu wird ebenfalls untersucht, ob die Referenz für den jeweiligen Cut-off-Wert eine Populationsnorm, die Stichprobe der jeweiligen Untersuchung oder sogar die Leistungsverteilung innerhalb von Subgruppen (z.B. auf Klassenebene) ist.

Methode

Unser Vorgehen orientiert sich an den bei Gough (2007) beschriebenen Schritten eines Systematic Reviews. Phase 1 dient nach Gough (2007) der Entwicklung der Fragestellungen. In Phase 2: Ein-

und Ausschlusskriterien wurden Studien in die Analyse eingeschlossen, welche die folgenden Kriterien erfüllten: 1) Es musste sich um einen empirischen Artikel in einer wissenschaftlichen Fachzeitschrift handeln. Diese Einschränkung erfolgte, um sicherzustellen, dass alle eingeschlossenen Beiträge ein Peer-Review-Verfahren durchlaufen haben. 2) Das Veröffentlichungsdatum musste zwischen 2000 und 2020 liegen. Die Beschränkung auf diesen Zeitraum erfolgte, um einen relativ aktuellen Überblick zu erhalten. 3) Die Stichprobe musste Schülerinnen und Schüler im Primar- oder Sekundarbereich umfassen. 4) Der Artikel musste mindestens eine Operationalisierung der Zielgruppe mit überdurchschnittlichen Schulleistungen (z. B. Leistungsstarke, high achievers, high performers) enthalten (z. B. durch die Stichprobenauswahl mit einem Fokus auf Leistungsstarke oder durch Bildung von Leistungsgruppen innerhalb einer größeren Stichprobe). 5) Es musste sich primär um Leistungen oder Kompetenzen in den Hauptfächern handeln ((fremd-)sprachliche, mathematische, naturwissenschaftliche Kompetenzen). Entsprechend wurden Studien, die sich eindeutig beispielsweise auf Hochleistungen im sportlichen oder musikalischen Bereich bezogen, ausgeschlossen. 6) Da die Operationalisierung von schulischer Leistungsstärke im engeren Sinne im Fokus stand, wurden Studien ausgeschlossen, die sich konzeptionell ausschließlich auf Begabte (z. B. durchgängige Nutzung der Begriffe „gifted“, „high cognitive ability“ statt „achievement“ oder „performance“) bezogen oder die ausschließlich Intelligenztestmaße, aber keine Maße zur Erfassung schulischer Leistungen heranzogen (z. B. Studien, die auf Basis der Study of Mathematically Precocious Youth, SMPY, entstanden sind, z. B. Lubinski & Benbow, 2006).

Phase 3: Recherchestrategie. Zunächst wurden deutschsprachige und internationale Datenbanken im Bereich Erziehungswissenschaft und Psychologie nach relevanten Studien durchsucht. Die Datenbanken waren im Einzelnen PsycArticles, Psyndex Literature & AV Media und ERIC. Da die unterschiedlichen Datenbanken unterschiedliche Thesauri für Schlagwörter und unterschiedliche Filtermöglichkeiten bieten, unterschieden sich die Suchbegriffe je nach Datenbank. Zusätzlich wurden weitere Artikel aufgenommen, die zum Beispiel von den Studien, die in den Datenbanken gefunden worden waren, zitiert wurden und die Einschlusskriterien dieses Systematic Review erfüllten. Die Suchen lieferten insgesamt 309 Ergebnisse.

Phase 4: Screening. Die Studien wurden in einem mehrschrittigen Verfahren gescreent und bei mangelnder Passung aussortiert. In einem ersten Schritt des Screenings wurde lediglich der Titel des Beitrags betrachtet. Auf diese Weise wurden Studien ausgeschlossen, bei denen bereits im Titel deutlich wurde, dass es sich nicht um Schülerinnen und Schüler im allgemeinbildenden Schulsystem handelte, dass es nicht um Schulleistungen in den Hauptfächern ging oder dass keine empirische Originalarbeit vorlag. In einem zweiten Schritt wurde dann das Abstract und in einem

dritten Schritt der Methodenteil des Artikels untersucht. 144 Artikel, welche die oben genannten Ein- und Ausschlusskriterien erfüllten, verblieben nach dem Screening von Abstract und Methodenteil in der weiteren Analyse.

Das Screening wurde für einen zufällig ausgewählten Anteil von 10% der Studien zusätzlich von einem zweiten Bewerter durchgeführt. Die Beobachterübereinstimmung betrug 83% (Cohens $\kappa = .66$). Unterschiede waren vor allem darin begründet, dass der Zweitbewerter das Kriterium 5 (Fachbezug) strenger auslegte und bei Notendurchschnitten ohne genauere Benennung der Fächer einen Fallausschluss empfahl, während die Erstautorin solche Studien in die Bewertung einschloss und nur solche Studien ausschloss, in denen es eindeutig nicht um Leistungen in Hauptfächern ging. Im Zweifelsfall wurden diese Studien zunächst eingeschlossen, da die Beurteilung der Qualität der Darstellung ein Teil der inhaltlichen Analyse werden sollte und auch das Wissen darüber, wie viele Studien eine unvollständige Darstellung der Operationalisierung von Leistungsstärke aufwiesen, von Interesse war.

Schließlich wurden die 144 verbleibenden Artikel danach bewertet, inwiefern sie ihren inhaltlichen Fokus tatsächlich auf leistungsstarke Schülerinnen und Schüler legten. Ein großer Teil der Studien kategorisierte Schülerinnen und Schüler zwar anhand ihrer Leistung und definierte in diesem Rahmen eine leistungsstarke (bzw. leistungsstärkere) Gruppe, allerdings waren leistungsstarke Schülerinnen und Schüler nicht ein Fokus der Untersuchung und die Gruppe wurde häufig sehr breit definiert (z.B. durch einen Median-Split). Vielmehr ging es in diesen Studien um die Robustheit oder Varianz der interessierenden Effekte (z.B. der Wirksamkeit von innerer und äußerer Differenzierung im Unterricht; Trautwein, Köller & Kämmerer, 2002) über das gesamte Leistungsspektrum hinweg. Diese Studien, die im Kern nicht Leistungsstärke untersuchten, wurden ausgeschlossen, so dass sich die Anzahl der Artikel auf die 55 Studien reduzierte, deren inhaltlicher Fokus die Leistungsstärke war (siehe Abbildung I-1). Ein Überblick über die Ergebnisse der Literaturrecherche und den Screeningprozess findet sich in den elektronischen Supplements ESM 1 und ESM 2.

Die eingeschlossenen Studien verteilten sich über eine Vielzahl an wissenschaftlichen Zeitschriften. Die häufigsten davon waren: High Ability Studies (N=6), Gifted Child Quarterly (N=4) und Journal of Educational Psychology (N=3). Die Publikationsjahre der Studien zeigen einen Anstieg an Studien ab 2012 mit einem Maximum im Jahr 2015.

Die meisten Studien stammten aus Nordamerika (N=27), gefolgt von Europa (N=16) und Asien (N=7). Drei Studien waren multinational und jeweils eine Studie stammte aus Afrika und Australien. In 10 der 55 Artikel wurden Daten deutscher Schülerinnen und Schüler verwendet.

Phase 6: Datenextraktion. Alle Studien wurden anhand der im Folgenden dargestellten Merkmale kodiert. Diese lassen sich in folgende Bereiche einteilen: 1) Merkmale, die die Operationalisierung beschreiben (Forschungsfragen 1 und 2) und 2) Merkmale, die forschungsmethodische und inhaltliche Aspekte der untersuchten Studien beschreiben. Zur Operationalisierung wurde kodiert, um welche Art von Indikator es sich handelte, auf welche Fächer sich die Leistungsstärke bezog, welcher Cut-off verwendet wurde, welche Bezugsnorm angelegt wurde, welche Stichprobengröße vorlag und welche Anteile Leistungsstarker an Stichprobe und Population sich aus den Angaben der Studie ableiten ließen. Forschungsmethodisch verfolgte der Großteil der Artikel einen quantitativen Ansatz, wobei neben den 47 quantitativ-beobachtenden Studien (72 %) auch 6 (quasi-)experimentelle Studien (9 %) vertreten waren. 11 Studien (17 %) waren qualitativ und eine verfolgte einen Mixed-Methods-Ansatz (2 %). Die Studien waren zu etwas weniger als der Hälfte (31 Studien, 48%) querschnittlich ausgerichtet. Die übrigen 23 Studien (35 %) waren Längsschnittstudien, 11 weitere Trendstudien (17 %). Die Stichprobengrößen schwankten sehr stark und lagen zwischen 9 und etwa 500,000 Teilnehmenden, bei einem Median von 616 Schülerinnen und Schülern. Der Großteil der Studien untersuchte Schülerinnen und Schüler in der Sekundarstufe I (39%). Grundschulen und Sekundarstufe II sowie Studien, die über mehrere

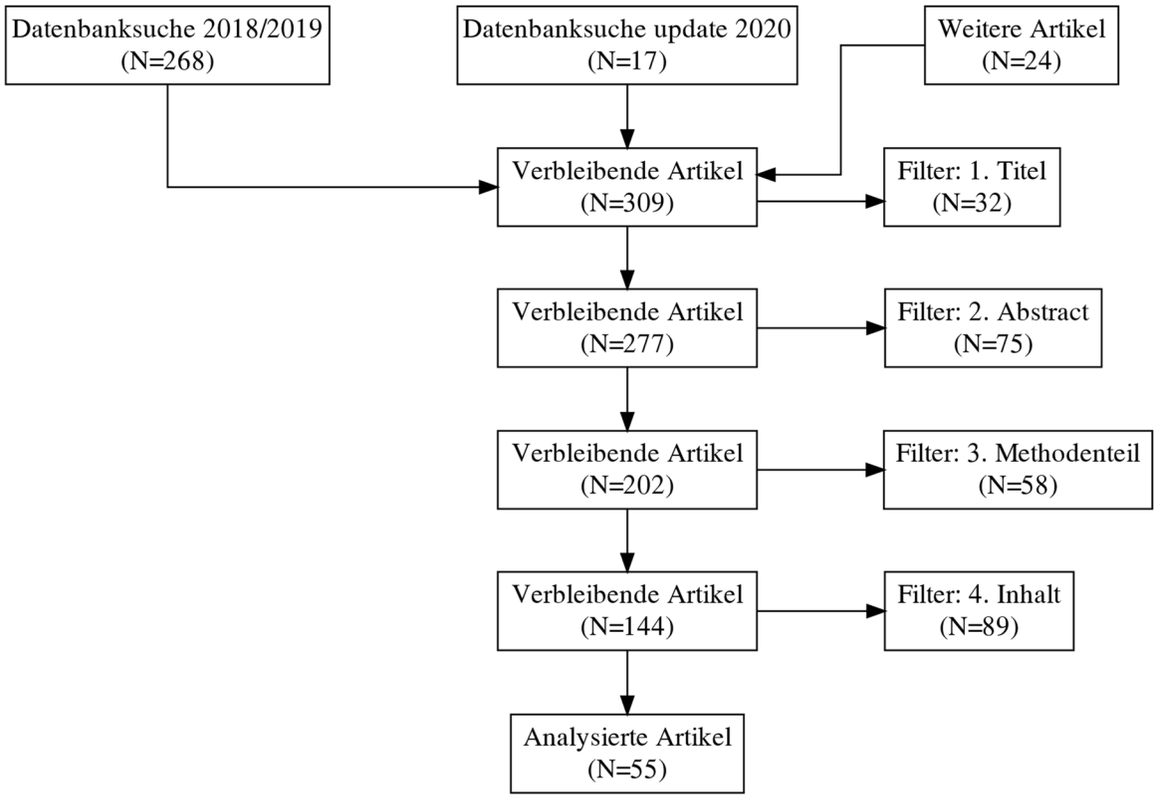


Abbildung I-1 Flowchart des Screening-Prozesses. Eine detaillierte Übersicht der Studien, die in den einzelnen Screeningschritten ausgeschlossen wurden findet sich im Elektronischen Supplement 1, der Code zur Reproduktion der Abbildung findet sich unter https://osf.io/jzkv6/

Tabelle I-1.

Verteilung der Studien auf die Altersgruppen bzw. Schulstufen

ISCED-Level	Häufigkeit	Anteil
Level 1: Grundschule	13	20%
Level 1-2: Längsschnitt Grundschule - Sekundarstufe I	4	6%
Level 2: Sekundarstufe I	25	39%
Level 1-3: Längsschnitt Grundschule - Sekundarstufe II	2	3%
Level 2-3: Längsschnitt Sekundarstufe I - II	7	11%
Level 3: Sekundarstufe II	14	22%

Anmerkung. ISCED = International Standard Classification of Education. Die Kategorien wurden gebildet, indem je nach Verfügbarkeit bei den Studien die Angabe zur Altersgruppe, zur Klassenstufe oder zur Schulform auf das entsprechende ISCED-Level transformiert wurde.

Bildungsstufen hinweg reichten, waren in etwa gleich vertreten mit jeweils rund 20 Prozent der Studien (s. Tabelle I-1). Eine Übersicht über die Studien inklusive der Kodierungen findet sich im Anhang I-A, Tabelle I-A2.

Acht der 55 Artikel enthielten eine oder mehrere Teil-Studien, wobei dies entweder eine zweite Stichprobe war, die untersucht wurde, oder unterschiedliche Operationalisierungen für Leistungsstärke an einer Stichprobe angewendet wurden. Diese unterschiedlichen Operationalisierungen oder Stichproben gingen als separate Studien in die Analysen ein, weshalb die Grundgesamtheit an Studien in der Analysestichprobe auf 65 stieg.

Ergebnisse

Forschungsfrage 1: Welche Indikatoren wurden zur Operationalisierung von Leistungsstärke herangezogen?

Die häufigsten Indikatoren von Leistungsstärke, die in den Studien Verwendung fanden, waren (1) Tests, (2) Noten und (3) die Zugehörigkeit zu einer speziellen Schule (z. B. magnet school, Schulen speziell für Leistungsstarke mit kompetitivem Auswahlverfahren), einer Schulform (z. B. Gymnasium) oder zu einem bestimmten Kurs für Leistungsstarke (z. B. Honors Kurse oder Advanced Placement Programs). Die letzten drei Kategorien wurden unter der Bezeichnung „Schulkontext“ zusammengefasst. In 57 Studien (88 %) wurde lediglich ein einzelner Indikatortyp zugrunde gelegt (Frage 1.b), wobei aber durchaus auch Werte desselben Typs miteinander verrechnet werden konnten (z. B. Notenschnitt über mehrere Fächernoten). In 8 Studien (12 %) wurden hingegen mehrere Indikatortypen kombiniert (s. Tabelle I-2). Beispielsweise beschrieben einige der Studien, welche die Zulassung zu einer speziellen Schule zugrunde legten, Zulassungsverfahren mit multiplen Kriterien (z. B. Noten, Motivationsschreiben, Intelligenztests oder Empfehlungsschreiben).

Tabelle I-2.

Indikatoren zur Operationalisierung der Zielgruppe der Leistungsstarken

Indikator	Häufigkeit	Anteil
Test	30	46%
Noten	22	34%
Schulkontext	21	32%
Einschätzungen	2	3%
Andere	2	3%
Kombination mehrerer Indikatoren	8	12%

Anmerkung. $N = 65$

In einer weiteren Analyse wurde der Frage nachgegangen, ob Leistungsstärke in den Studien fachbezogen oder als fächerübergreifend erfasst wurde (Frage 1.c). Von den untersuchten Studien bezogen sich 45 % in ihrer Definition von Leistungsstärke auf Leistungen in einem Fach, 39 % legten die Leistungen in mehreren Fächern zugrunde. Dabei verwendeten 16 Studien einen Durchschnittswert. In neun Studien wurde Leistungsstärke ebenfalls über mehrere Fächer bestimmt, aber kein Notendurchschnitt verwendet, sondern andere Verfahren wie z. B. Profilanalysen angewandt (s. Abbildung I-2).

Forschungsfrage 2: Welche Cut-off-Werte wurden berichtet?

Auch bei einer Verwendung vergleichbarer Indikatoren kann jedoch die Übertragbarkeit von Ergebnissen zwischen verschiedenen Studien fraglich sein, wenn sich die festgelegten Cut-off-Werte zur Bildung von Leistungsgruppen stark unterscheiden. Darüber hinaus kann ein ähnlicher absoluter Cut-off-Wert eine sehr unterschiedliche Bedeutung haben, wenn eine soziale Referenznorm verwendet wird und unterschiedliche Vergleichsgruppen dafür vorliegen. Im Folgenden wird daher die Untersuchung der genutzten Cut-off-Werte gemeinsam mit der Betrachtung der angelegten Vergleichsnormen diskutiert.

Bei quantitativen Indikatoren können Cut-off-Werte zwischen verschiedenen Studien verglichen werden, nachdem sie auf eine gemeinsame Skala transformiert wurden. Bei Testwerten geschieht dies häufig durch Perzentilwerte, die auf der Werteverteilung in der Normstichprobe beruhen. Dies ist allerdings nur bei normierten Tests oder aber in Erhebungen möglich, in denen die Verteilung in der Stichprobe der Verteilung in der Population entspricht. Von den 30 Studien, die Tests eingesetzt haben, haben 27 (90 %) normierte Leistungstests verwendet. Von diesen nutzten 19 Studien Daten aus Large-Scale-Assessment-Studien oder administrative Daten aus staatlichen Lernstandserhebungen. Sieben dieser Studien verwendeten die Angabe der erreichten Kompetenzstufe (zur Entwicklung von Kompetenzstufenmodellen, siehe z.B. Pant, Tiffin-Richards & Köller, 2010) und damit eine kriteriale Bezugsnorm zur Definition von

Leistungsstärke. Die übrigen 12 Studien legten eine soziale Bezugsnorm an. Die Cut-off-Werte dieser Studien lagen zwischen 4 und 50 Prozent (Mdn = 25 %)

Acht Studien (29.6 %) setzten normierte Tests ein, wie zum Beispiel den Iowa Test of Basic Skills (Hoover, Dunbar & Frisbie, 2001) oder den Woodcock-Johnson Test of Achievement (Woodcock, McGrew & Mather, 2001). Fünf dieser Studien verwendeten die Populationsnormen bei der Beschreibung ihrer Stichprobe der leistungsstarken Schülerinnen und Schüler. Die entsprechenden Cut-off-Werte lagen zwischen 2 und 15 Prozent (Mdn = 4 %) und somit deutlich geringer als bei den Studien auf Basis von Large-Scale-Assessments. Drei Studien wählten alternative Identifikationsverfahren: Die eine Studie definierte die besten zwei Prozent auf Klassenebene, die jedoch nicht gleichzeitig über dem 98. Perzentil der Normstrichprobe liegen durften, als leistungsstark (für eine Erläuterung s. Rambo-Hernandez & McCoach, 2015) und die anderen beiden Studien nutzten statistische Verfahren, um die Gruppe der Leistungsstarken zu definieren (Latente Profilanalyse: Wang, Eccles & Kenny, 2013, Faktorenanalyse: Robinson, Lanzi, Weinberg, Ramey & Ramey, 2002).

Zusammenfassend lässt sich sagen, dass die Cut-off-Werte bei Tests erheblich streuten. Dort, wo sie auf eine gemeinsame Skala transferierbar waren, nahmen sie Werte zwischen 2 und 50 Prozent an (Mdn = 10 %). Bezogen auf die Stichproben der Studien, wurde im Mittel etwa ein Siebtel der Schülerinnen und Schüler mithilfe von Testwerten als leistungsstark definiert (M = 15 %, SD = 12 %).

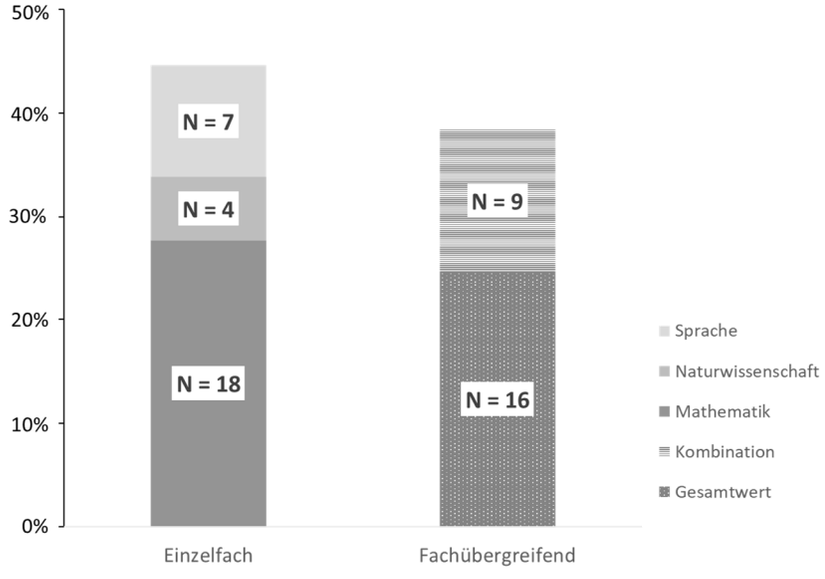


Abbildung I-4. Fachbezug der Definitionen von Leistungsstärke. N = 11 Studien machten keine Angaben bzw. nutzten keine Fachleistungen zur Operationalisierung. Mit Gesamtwert sind Studien gemeint, die Notenschnitte bildeten oder einen präexistenten Schnitt (z. B. Abiturgesamtpunktzahl) nutzten. Mit Kombination sind Studien gemeint, die Fachleistungen anders zusammenfassten, z.B. durch Faktorenanalysen, Latente Profilanalysen, und- bzw. oder-Kriterien. Der Code zur Reproduktion der Abbildung findet sich unter https://osf.io/jzkv6/.

Von den 22 Studien (34 %), die Noten nutzten, um die Gruppe der Leistungsstarken zu identifizieren, nutzten 16 Studien einen konkreten Notenwert und damit im Prinzip einen kriterialen Cut-off-Score, 3 Studien gaben einen Anteil leistungsstarker Schülerinnen und Schüler vor (2 %: Forgasz & Hill, 2013;Vock, Köller & Nagy, 2013, bzw. 10 %: Sontag & Stoeger, 2015). In einer Studie (Mourgues, Hein, Tan, Diffley III & Grigorenko, 2016) wurde eine latente Klassenanalyse durchgeführt, um über das höchstgelegene Leistungsprofil Leistungsstärke zu bestimmen. Zwei weitere Studien machten keine Angaben zum Cut-off-Wert. Die eingesetzten Notenskalen waren notwendigerweise divers und schwer vergleichbar, da die Stichproben aus unterschiedlichen Ländern stammten (s. Tabelle I-A3, Anhang I-A). Forschende versuchten, diese Schwierigkeit zu umgehen, indem sie das jeweilige Bewertungssystem möglichst genau beschrieben, indem sie eine Umrechnung in den in den USA geläufigen GPA (Grade Point Average) vornahmen oder indem der Anteil an Schülerinnen und Schülern angegeben wurde, die diese Noten typischerweise erhalten (z. B. Salmela & Uusiautti, 2015). Von den 14 Studien (22 %), die ausschließlich Noten zur Einteilung der Leistungsstarken verwendeten, bestand die Stichprobe in fünf Studien komplett aus Leistungsstarken (die Auswahl erfolgte also direkt bei der Stichprobenziehung). Bei weiteren acht Studien lagen die Stichprobenanteile Leistungsstarker zwischen 3 und 50 Prozent (M = 29 %). Studien, die neben Noten weitere Kriterien verwendeten, legten einen geringeren Cut-off-Wert bei den Noten an.

Bei der Operationalisierung über den Schulkontext ist in der Regel Hintergrundwissen über diesen vonnöten, um die Exklusivität der Gruppe beurteilen zu können. In manchen Studien (N = 12) wurden Kinder und Jugendliche an speziellen Schulen für Leistungsstarke untersucht. Hier ist eine Vergleichbarkeit zu anderen Studien schwierig, da häufig idiosynkratische Fallbeschreibungen die Datengrundlage bildeten. Teils wurde angegeben, wie umfangreich das Auswahlverfahren einer Schule ist oder welche Leistungen zu erbringen sind, um zugelassen zu werden. In manchen Studien werden auch Auswahlquoten angegeben. In lediglich einem kleinen Teil der Studien geht es dabei allerdings um Fragestellungen, die ganz explizit eine konkrete Schule oder Schulform betreffen.

Diskussion

Vollständigkeit der Angaben zur Operationalisierung

Das vorliegende Review fragte danach, wie in den Studien der letzten 20 Jahre Leistungsstärke operationalisiert wurde. Ausgewählt wurden insbesondere die Studien, deren erklärtes Ziel es war, leistungsstarke Schülerinnen und Schüler in schulischen Hauptdomänen zu untersuchen. Die Studien zeichneten sich durch eine große Diversität in ihren methodischen Merkmalen aus. Sie

zeigten eine große Streuung in Stichprobenumfang, Forschungs- und Erhebungsdesign und untersuchten verschiedene Altersgruppen im allgemeinbildenden Schulsystem.

Der Detailgrad der Beschreibungen der Leistungsstärke unterschied sich zwischen den Studien sehr stark. Teilweise waren beinahe alle relevanten Informationen bereits im Abstract enthalten, teilweise im Methodenteil vorhanden, teilweise musste jedoch im gesamten Artikel nach den relevanten Informationen gesucht werden. Ziegler und Raul (2000) fanden bei 11 % der Studien keine Angaben zur Identifikation Begabter, während in den hier untersuchten Studien nur vereinzelt relevante Informationen gänzlich fehlten. Zwei Studien machten keine Angaben zum gewählten Cut-off in der Stichprobe, drei Studien machten keine Angabe dazu, welche Fächer zur Bestimmung der Leistungsstärke betrachtet wurden, bei sieben Studien fehlten Angaben zur Stichprobengröße und in fünf Studien wurden keine Angaben zum Anteil Leistungsstarker an der Stichprobe gemacht. Während einige der fehlenden Angaben lediglich die Interpretierbarkeit des Begriffs von Leistungsstärke erschweren, der zugrunde lag (z. B. was bedeutet eine bestimmte Note? Welchem Anteil an der Population entspricht ein bestimmter Cut-off?), ist das komplette Fehlen von Angaben, die die Identifikation nachvollziehbar machen, eine Ausnahme und trat bei lediglich zwei Studien auf. Bei diesen Studien wurde lediglich berichtet, dass Leistungsstarke aufgrund ihrer Schulleistungen identifiziert wurden (Adeloun et al., 2015; Mioduser & Betzner, 2009).

Diskussion zentraler Ergebnisse

Forschungsfrage 1: Welche Indikatoren werden zur Operationalisierung von Leistungsstärke herangezogen?

Im vorliegenden Review zu leistungsstarken Schülerinnen und Schülern wurden als Indikatoren für Leistungsstärke am häufigsten Tests und Noten herangezogen, die auch in der Übersichtsarbeit von Carman (2013) zur Klassifikation von Begabten eine wichtige Rolle gespielt hatten. Die Vorauswahl Leistungsstarker durch das Bildungssystem, beispielsweise durch Selektionsprozesse in segregierten Bildungssystemen, war in der vorliegenden Übersichtsarbeit beinahe ebenso häufig wie die Identifikation Leistungsstarker anhand von Noten. Bei Ziegler und Raul (2000) nutzte die Mehrheit der Studien multiple Kriterien zur Identifikation Begabter; im vorliegenden Beitrag fand eine Kombination mehrerer Indikatoren hingegen lediglich bei einem kleinen Teil der Studien statt. Zusammenfassend lässt sich schlussfolgern, dass leistungsstarke Schülerinnen und Schüler eine Gruppe darstellen, die sich zwar zunächst konzeptionell von begabten (im Sinne von hochintelligenten) Schülerinnen und Schülern unterscheiden lässt: Bei Begabung liegt die Betonung stärker auf einem Potenzial, welches noch nicht ausgeschöpft sein muss, während bei

Leistungsstärke die Performanz und damit bereits gezeigte Leistungen im Mittelpunkt stehen. Bei deren Identifikation in der Forschungspraxis gibt es aber durchaus große Überschneidungen. So wäre es basierend auf den Ergebnissen der Reviews in vielen Fällen schwierig, allein von der Operationalisierung einer beforschten Gruppe darauf zu schließen, ob Forschende hier begabte oder leistungsstarke Schülerinnen und Schüler untersuchen wollten. Die Entwicklungen hin zu einer stärkeren Integration von Begabung und Leistung im Rahmen neuerer Talententwicklungsmodelle (Preckel et al., 2020; Subotnik et al., 2011) scheinen vor diesem Hintergrund folgerichtig.

Eine weitere Hauptfrage in diesem Zusammenhang war die Domänenspezifität von Leistungsstärke. Es fanden sich sowohl Studien, die sich auf Leistungen in einem einzelnen Fach bezogen als auch Studien, die mehrere Fächer in ihre Definition von Leistungsstärke einbezogen. Das Spannungsfeld zwischen der fachspezifischen Manifestation von Leistung und dem fächerübergreifenden Konzept der „Einserschülerin“ bzw. des „Einserschülers“ äußert sich auch darin, dass einige Studien die Analysen für jedes Fach separat wiederholten, aber auch die Bildung von Leistungsprofilen über mehrere Fächer hinweg keine Seltenheit war. Zusätzliche Auswertungen zeigten, dass an Grundschulen häufiger Notenschnitte und in der Sekundarstufe I häufiger Einzelfächer oder Fächerprofile zur Klassifikation Leistungsstarker genutzt wurden.

Legt man aktuelle Begabungsmodelle zugrunde, so passen beide Beobachtungen—sowohl die Überschneidung der beiden Konstrukte Leistungsstärke und Begabung, als auch die fachspezifische und fachübergreifende Betrachtung von Leistungsstärke—in den Talententwicklungsansatz aktueller Begabungsmodelle (Preckel et al., 2020; Subotnik et al., 2011), welche Begabung als Entwicklung von einer breit angelegten (kognitiven) Potenz hin zur Ausbildung immer spezialisierterer Kompetenzen und Fähigkeiten begreifen.

Forschungsfrage 2: Welche Cut-off-Werte werden verwendet?

Bei der Definition eines angemessenen Cut-off-Wertes zur Abgrenzung der leistungsstarken Schülerinnen und Schüler auf dem Leistungsspektrum zeigte sich am deutlichsten, was Forschende unter Leistungsstärke verstehen. Dabei muss unterschieden werden zwischen der Festlegung eines Anteils Leistungsstarker an der Stichprobe und einem Cut-off-Wert, der einen Rückschluss auf die Anteile in der Population und damit eine Vergleichbarkeit über Studien hinweg zulässt. Dass dies unterschiedliche Aspekte sind, zeigt sich beispielsweise daran, dass viele Studien deutliche Abweichungen zwischen dem Stichprobenanteil Leistungsstarker und dem angenommenen Populationsanteil Leistungsstarker aufwiesen. Um diesen Populationsparameter feststellen zu können, müssen detaillierte Informationen zu den eingesetzten Tests und der Stichprobe vorliegen.

Eine Übertragbarkeit des Stichprobenanteils Leistungsstarker auf die Verhältnisse in der Population wird nur dann möglich, wenn Large-Scale-Assessments (ggf. mit Populationsgewichten), Vollerhebungen, eine dezidierte Zufallsauswahl oder normierte Tests vorliegen und eine Populationsnorm als Vergleichsmaßstab zur Verfügung steht. Bei Studien, die Noten als Leistungsindikator nutzen, stellte sich dieses Problem in besonderem Ausmaß, da Noten, wie eingangs erwähnt, kaum zwischen verschiedenen Klassen, noch weniger aber häufig zwischen unterschiedlichen Ländern vergleichbar sind.

Bei den Studien, bei denen dieser Rückschluss auf die Population möglich war, lagen die Cut-off-Werte generell strenger. Die Hälfte dieser Studien legte einen Cut-off-Wert an, der weniger als 10 % der Population als leistungsstark definierte. Etwa ein Sechstel der Studien definierte zwei Prozent der Population als leistungsstark – es kann vermutet werden, dass Forschende sich mit diesem strengen Kriterium an der vorherrschenden Definition von Hochbegabung orientierten, die ca. zwei Prozent der Population (IQ-Wert von mindestens 130 Punkten) umfasst (Rost & Buch, 2018). Bei diesen Studien könnte man bereits fragen, inwiefern ein konzeptioneller Unterschied zur Begabung besteht.

Allerdings machte die Kodierung des Cut-off-Kriteriums die größten Schwierigkeiten. Forschende machten häufig unvollständige Angaben, die eine Einschätzung des intendierten (d. h. populationsbezogenen) Cut-off-Wertes verhinderten. Über die Hälfte der Studien ließ keinen Schluss auf das intendierte Cut-off-Kriterium zu. Die Ursache hierfür war meist, dass vorausgelesene Stichproben verwendet und idiosynkratische Indikatoren genutzt wurden, die sehr spezifisch für die jeweilige Stichprobe konstruiert wurden. Teilweise entsteht der Eindruck, teilweise wird explizit gesagt, dass nicht aufgrund inhaltlicher, sondern aufgrund statistischer Überlegungen (Schätzbarkeit mittels parametrischer Verfahren, vergleichbare Gruppengrößen) ein bestimmter Cut-off-Wert gewählt worden war. Teils wurden auch unterschiedliche Cut-off-Werte nebeneinandergestellt, um empirisch den Einfluss der Entscheidung für einen Wert auf die Ergebnisse zu prüfen (Gagné, 2016; Möller & Pohlmann, 2010; Neuendorf et al., 2020; Rambo-Hernandez & McCoach, 2015; Schurtz, Pfost & Artelt, 2014; Zhou, Fan, Wei & Tai, 2017).

Einschränkungen der Generalisierbarkeit

Die Ergebnisse unserer Studie beruhen notwendigerweise auf einem spezifischen Ausschnitt der wissenschaftlichen Literatur. Dieser ist zunächst durch den Untersuchungszeitraum eingeschränkt. Es ist gut möglich, dass die Operationalisierung von Leistungsstärke in der Vergangenheit anders aussah, insbesondere auf Grundlage der sich ändernden Rahmenbedingungen der Forschung (z.B. Verfügbarkeit von Large-Scale-Assessment-Daten, Weiterentwicklung von Theorien und

Forschungsparadigmen) und in der Zukunft anders aussehen wird. Ziel unserer Studie war aber die Abbildung des aktuellen Verständnisses des Begriffs Leistungsstärke auch vor dem Hintergrund aktueller bildungspolitischer und fachlicher Diskussionen. Der von uns untersuchte Ausschnitt ist auch stark von den genutzten Fachdatenbanken und Suchbegriffen geprägt. So nutzten wir als einzige deutschsprachige Datenbank Psyndex, entschieden uns aber gegen die Nutzung von FIS-Bildung als deutschsprachige Fachdatenbank im Bereich der Pädagogik. Grund hierfür war in erster Linie die fehlende Möglichkeit bei FIS-Bildung, nach Beiträgen in wissenschaftlichen Fachzeitschriften zu filtern. Weiterhin war der Begriff der Leistungsstärke nicht in der Schlagwortliste von FIS-Bildung enthalten. Wir bemühten uns, durch Variation der Suchbegriffe (top student, high performer, high achiever) eine größere Menge passender Artikel zu identifizieren und entschieden uns schließlich datenbankspezifisch Begriffe zu verwenden, die auf Basis erster Sichtungen der Ergebnisse ein sinnvolles Spektrum an Resultaten ergaben. In einigen Punkten ist die Entscheidung über einen Aus- oder Einschluss einer Studie zu einem gewissen Grad subjektiv – das traf insbesondere auf die Frage zu, welche Studien einen hinreichenden Fokus auf leistungsstarke Schülerinnen und Schüler aufwiesen. Wir bemühten uns daher, möglichst umfassend den Rechercheprozess zu dokumentieren, um unser Vorgehen nachvollziehbar und prüfbar zu machen (s. Tabelle I-A1 im Anhang I-A, elektronische Supplements ESM1 und ESM2, sowie weitere Materialien im Open Science Framework (https://osf.io/jzkv6/).

Schlussfolgerungen

Für die Forschung, die sich mit leistungsstarken Schülerinnen und Schülern befasst, ergeben sich vier zentrale Empfehlungen. Das Systematic Review hat verdeutlicht, dass leistungsstarke Schülerinnen und Schüler aus unterschiedlichsten theoretischen Perspektiven und mit unterschiedlichen Zielstellungen untersucht werden. Daraus folgt, dass es erstens wichtig ist, zu verdeutlichen, welcher theoretische Begriff von Leistungsstärke bei der Untersuchung im Vordergrund steht. Dieser theoretische Hintergrund sollte die Grundlage der Definition leistungsstarker Schülerinnen und Schüler darstellen und eine Begründung für den Operationalisierungsansatz bilden. Ist beispielsweise der bedeutsame Aspekt für eine spezifische Fragestellung die soziale Konstruktion von Leistungsstärke (zum Beispiel Peer-Interaktionen oder Viktimisierung Leistungsstarker), dann wäre ein niedrig inferentes und klar erkennbares Merkmal, welches Kinder und Jugendliche gegenüber ihren Klassenkameraden offensichtlich als leistungsstark auszeichnet, wie zum Beispiel die Schulnoten, eine soziale Bezugsnorm, und ggf. ein fächerübergreifender Ansatz angemessen. Wenn das Thema der Studie hingegen in der Wirksamkeit bestimmter Unterrichtsmethoden liegt, wäre zu erwarten, dass eher ein

fachspezifischer, objektiver und über Klassen hinweg vergleichbarer Testwert als Leistungsindikator genutzt wird. Liegt der Fokus eher auf dem Erreichen von Kompetenzen in einem bestimmten Fach, dann wäre ein Kompetenztest und möglicherweise eine kriteriale Bezugsnorm die vorzuziehende Wahl. Ebenso sollte, wenn das TAD-Modell (Preckel et al., 2020) die theoretische Grundlage einer Untersuchung bildet, die Definition je nach Phase im Talententwicklungsprozess domänenübergreifend (frühe Phasen) oder fachspezifisch sein (fortgeschrittene Phasen). Hierbei ist auch zu beachten, dass bestimmte inhaltliche Fragestellungen an Untersuchungsdesigns gebunden sind, welche wiederum die Art der Operationalisierung von Leistungsstärke mit bedingen. So benötigen beispielsweise Fragestellungen, die sich mit mikrosoziologischen Prozessen befassen, häufiger einen qualitativen Zugang, der wiederum spezielle Rahmenbedingungen (z.B. reichhaltige Kontextinformationen, Möglichkeit zum Einbezug multipler Perspektiven/Intersubjektivität) für die Definition und Operationalisierung von Leistungsstärke schafft. Bei anderen Fragestellungen (wie zum Beispiel zu gesellschaftlichen Disparitäten) sind repräsentative, teils länderübergreifende Stichproben essenziell, welche sich insbesondere in Large-Scale-Assessment Studien finden lassen. In den Large-Scale-Assessments liegen fachspezifische Testwerte vor, die auch im hohen Leistungsbereich gut differenzieren. Daher und aufgrund der großen Stichprobenumfänge ist die Möglichkeit gegeben, auch kleinere Gruppen leistungsstarker Schülerinnen und Schüler auf belastbare Weise zu identifizieren.

Obgleich eine gemeinsame Definition des Begriffs bisher nicht existiert und eine Vereinheitlichung über alle Studien hinweg kaum realistisch ist, wäre ein geteiltes Verständnis des Konstrukts Leistungsstärke zumindest innerhalb von Forschungsfeldern ein Fortschritt und würde dafür sorgen, dass Forscherinnen und Forscher besser auf vorangegangenen Forschungsarbeiten aufbauen können. Besonders die Forschungsfelder, die sich mit der Entwicklung und Förderung von Kindern und Jugendlichen mit herausragenden akademischen Leistungen befassen, wie zum Beispiel die neuere Begabungsforschung, würden von einer einheitlichen Definition von Leistungsstärke profitieren. Angesichts des stärker fachspezifischen Ansatzes der Talentförderung in neueren Begabungsmodellen (Preckel et al., 2020) wäre eine Definition, die eine fachspezifische Operationalisierung erlaubt, aber gleichzeitig Vergleiche zwischen Leistungsstarken unterschiedlicher Talentdomänen ermöglicht, vorzuziehen. Zu diesem Zweck erscheint eine Definition anhand des Populationsanteils (ähnlich wie die Konvention der besten 2 % in der Begabungsforschung) sinnvoll. Wie groß dieser Anteil bemessen sein sollte, ist eine Frage, die noch im Forschungsfeld auszuhandeln ist. Dabei sollte idealerweise die praktische Relevanz der Definition bedacht werden. Bisher kursieren einerseits operationale Definitionen, die sich an der Größe der Begabtengruppe orientieren und somit einen sehr kleinen Anteil akademisch

leistungsstarker Kinder und Jugendlicher annehmen (Rost, 2009), aber auch Begründungen, die zum Beispiel auf die Kompetenzstufendefinitionen von Large-Scale-Assessments Bezug nehmen und einen größeren Anteil der Schülerinnen und Schüler beschreiben (z. B. Köller & Baumert, 2017; Neuendorf et al., 2020).

Unabhängig von der theoretischen Begründung für die Gruppendefinition sollten jedoch zweitens, wenn leistungsstarke Kinder und Jugendliche als eigene Gruppe dargestellt werden, die Stichproben- und Instrumentenbeschreibung im Methodenteil so konkret sein, dass Leserinnen und Leser einschätzen können, wie die beschriebene Gruppe definiert worden ist. Falls eine vorausgelesene Stichprobe genutzt wird, sollte erkennbar sein, auf welchen Anteil der Population sich die Aussagen über Leistungsstarke generalisieren lassen (siehe auch Simons, Shoda & Lindsay, 2017). Werden kriteriale Normen genutzt, dann sollte expliziert werden, welche Bedeutung dieses Kriterium besitzt und möglichst auch, welcher Anteil der Population dieses Kriterium typischerweise erreicht.

Drittens sollte in Forschungsarbeiten diskutiert werden, aus welchen Gründen genau diese Operationalisierung gewählt wurde (z. B. theoretische Überlegungen, statistische Notwendigkeiten). Da Schulleistung ein kontinuierliches Merkmal ist, stellt sich auch die Frage, in welchem Fall eine Gruppeneinteilung überhaupt notwendig ist (z. B. wenn es um die Qualifikation zur Teilnahme an bestimmten Förderangeboten geht) und in welchen Fällen auch das gesamte Leistungsspektrum genutzt werden könnte, um Fragestellungen mit Bezug zu leistungsstarken Schülerinnen und Schülern zu beantworten.

Oftmals lässt sich die Operationalisierung von Leistungsstärke nicht eindeutig aus Theorie und Zielstellung ableiten. Es liegt daher an den Forschenden, die gewählte Operationalisierung zu begründen. Daher wird hier viertens und abschließend eine Möglichkeit vorgestellt, die Stabilität von Ergebnissen für verschiedene Operationalisierungen von leistungsstarken Kindern und Jugendlichen zu prüfen. Die Multiversumsanalyse (Steegen, Tuerlinckx, Gelman & Vanpaemel, 2016) ist eine Methode, bei der systematisch verschiedene Operationalisierungen gegenübergestellt werden. Dabei werden Entscheidungen während der Datenaufbereitung systematisch variiert und die Effekte dieser unterschiedlichen Spezifikationen miteinander verglichen. So kann beurteilt werden, ob die Ergebnisse diesen Variationen standhalten. So könnten bei Studien zu leistungsstarken Schülerinnen und Schülern Ergebnisse für verschiedene Konfigurationen nachvollziehbarer Entscheidungen bezüglich der verwendeten Indikatoren, Cut-off-Kriterien, Vergleichsmaßstäbe und Fachbezüge verglichen und die Varianz der Ergebnisse in Abhängigkeit von diesen Parametern abgeschätzt werden. Eine Möglichkeit, die Ergebnisse

solcher systematisch permutierten Kombinationen darzustellen und die Effekte visuell zu vergleichen, besteht in der Spezifikationskurve (Simonsohn, Simmons & Nelson, 2020). Diese Darstellung ermöglicht es herauszufinden, welche Entscheidung den größten Anteil an Variation in den Ergebnissen erklärt (z. B. Neuendorf et al., 2020).

Ausblick

Was Preckel et al. (2020) über die Begabungsforschung schreiben, trifft auch auf die Forschung zu Leistungsstärke im schulischen Kontext zu: Das Feld würde von mehr Koordination zwischen Forschenden und einer stärkeren Verzahnung von Theorie und Praxis profitieren. Leistungsstarke Schülerinnen und Schüler werden in verschiedenen Teilgebieten der Psychologie, wie differenzieller Psychologie, Motivationspsychologie, kognitiver Psychologie, pädagogischer Psychologie und auch Begabungsforschung untersucht. Die vorliegende Übersichtsarbeit benennt mit der transparenten Operationalisierung von Leistungsstärke eine Bedingung für die Beurteilung der Übertragbarkeit von Ergebnissen zwischen den Fachbereichen aber auch zwischen verschiedenen Studien innerhalb von Forschungsgebieten. Dadurch wird die Integration von Wissen aus den verschiedenen Teilbereichen ermöglicht. Darüber hinaus soll die Relevanz der Forschungsergebnisse für Praxisinitiativen, wie dem eingangs benannten LemaS-Projekt, erhöht werden. Auch hier kann die vorliegende Arbeit einen Beitrag leisten, wenn aufbauend auf ihren Resultaten Indikatoren entwickelt werden, anhand derer die Ergebnisse von Förderbemühungen beurteilt werden können.

Literaturverzeichnis

Ein vorangestellter Stern kennzeichnet Artikel, welche in die Analysen des systematischen Reviews eingegangen sind.

Acar, S., Sen, S. & Cayirdag, N. (2016). Consistency of the performance and nonperformance methods in gifted identification. *Gifted Child Quarterly*, *60*(2), 81–101. https://doi.org/10.1177/0016986216634438

*Adelodun, G. A. & Asiru, A. B. (2015). Instructional resources as determinants of english language performance of secondary school high-achieving students in Ibadan, Oyo State. *Journal of Education and Practice*, *6*(21), 195–200.

*Allen, W. & Griffin, K. (2006). Mo' Money, Mo' Problems? High-achieving black high school students' experiences with resources, racial climate, and resilience. *Journal of Negro Education, 75*(3), 478–494. https://www.jstor.org/stable/40026816

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (Springer-Lehrbuch, 4., vollst. überarb. u. erw. Aufl.). Berlin, Heidelberg: Springer Medizin. https://doi.org/10.1007/3-540-28507-5

*Assouline, S. G., Ihrig, L. M. & Mahatmya, D. (2017). Closing the excellence gap: Investigation of an expanded talent search model for student selection into an extracurricular STEM program in rural middle schools. *Gifted Child Quarterly, 61*(3), 250–261. https://doi.org/10.1177/0016986217701833

*Barron, B. (2000). Problem solving in video-based microworlds: Collaborative and individual outcomes of high-achieving sixth-grade students. *Journal of Educational Psychology, 92*(2), 391–398. https://doi.org/10.1037/0022-0663.92.2.391

*Bergold, S., Kasper, D., Wendt, H. & Steinmayr, R. (2020). Being bullied at school: the case of high-achieving boys. *Social Psychology of Education*, *23*(2), 315–338. https://doi.org/10.1007/s11218-019-09539-w

*Bergold, S., Wendt, H., Kasper, D. & Steinmayr, R. (2017). Academic competencies. Their interrelatedness and gender differences at their high end. *Journal of Educational Psychology*, *109*(3), 439–449. https://doi.org/10.1037/edu0000140

*Berkowitz, E. & Cicchelli, T. (2004). Metacognitive strategy use in reading of gifted high achieving and gifted underachieving middle school students in New York City. *Education and Urban Society*, *37*(1), 37–57. https://doi.org/10.1177/0013124504268072

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D. et al. (2012). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 227–259). Münster [u.a.]: Waxmann.

Brühwiler, C. & Helmke, A. (2018). Determinanten der Schulleistung. In D. H. Rost, J. R. Sparfeldt & S. Buch (Hrsg.), *Handwörterbuch pädagogische Psychologie* (Beltz Psychologie 2018, 5., überarbeitete und erweiterte Auflage). Weinheim: Beltz.

Bundesministerium für Bildung und Forschung; Kultusministerkonferenz. (2016, 28. September). Gemeinsame Initiative von Bund und Ländern zur Förderung leistungsstarker und potenziell besonders leistungsfähiger Schülerinnen und Schüler.

*Bütüner, S. Ö. & Filiz, M. (2017). Exploring high-achieving sixth grade students' erroneous answers and misconceptions on the angle concept. *International Journal of Mathematical Education in Science and Technology, 48*(4), 533–554. https://doi.org/10.1080/0020739X.2016.1256444

Carman, C. A. (2013). Comparing apples and oranges. Fifteen years of definitions of giftedness in research. *Journal of Advanced Academics*, *24*(1), 52–70. https://doi.org/10.1177/1932202X12472602

*Carroll, P. E. & Bailey, A. L. (2016). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. Language Testing, 33(1), 23–52. https://doi.org/10.1177/0265532215576380

*Carter, D. J. (2008). Achievement as resistance: The development of a critical race achievement ideology among black achievers. *Harvard Educational Review, 78*(3), 466–497.

*Carter Andrews, D. J. (2012). Black achievers' experiences with racial spotlighting and ignoring in a predominantly white high school. *Teachers College Record, 114*(10), 1–46.

*Castejón, A. & Zancajo, A. (2015). Educational differentiation policies and the performance of disadvantaged students across OECD countries. *European Educational Research Journal*, *14*(3-4), 222–239. https://doi.org/10.1177/1474904115592489

*Cheung, K. C. (2017). The effects of resilience in learning variables on mathematical literacy performance: A study of learning characteristics of the academic resilient and advantaged low achievers in Shanghai, Singapore, Hong Kong, Taiwan and Korea. *Educational Psychology, 37*(8), 965–982. https://doi.org/10.1080/01443410.2016.1194372

*Clausen, M., Weingarten, J. & Wegner, H. (2013). Unterrichtsqualität an einer besonderen Schule: Videobasierte Evaluation eines Oberstufen-Internats für leistungsstarke und hoch motivierte Schülerinnen und Schüler. *Gruppendynamik und Organisationsberatung, 44*(3), 301–321. https://doi.org/10.1007/s11612-013-0218-y

*Crawford, C., Macmillan, L. & Vignoles, A. (2017). When and why do initially high-achieving poor children fall behind? *Oxford Review of Education*, *43*(1), 88–108. https://doi.org/10.1080/03054985.2016.1240672

*Ee, J., Moore, P. J. & Atputhasamy, L. (2003). High-achieving students: Their motivational goals, self-regulation and achievement and relationships to their teachers' goals and strategy-based instruction. *High Ability Studies*, *14*(1), 23–39. https:/doi.org/10.1080/13598130304094

*Erbas, A. K. & Bas, S. (2015). The contribution of personality traits, motivation, academic risk-taking and metacognition to the creative ability in mathematics. *Creativity Research Journal, 27*(4), 299–307. https://doi.org/10.1080/10400419.2015.1087235

*Flores-Gonzalez, N. (2005). Popularity versus respect: School structure, peer groups and latino academic achievement. international *Journal of Qualitative Studies in Education (QSE), 18*(5), 625–642.

*Forgasz, H. J. & Hill, J. C. (2013). Factors implicated in high mathematics achievement. *International Journal of Science and Mathematics Education*, *11*(2), 481–499. https://doi.org/10.1007/s10763-012-9348-x

*Gagné, F. (2016). From noncompetence to exceptional talent. Exploring the Range of Academic Achievement Within and Between Grade Levels. *Gifted Child Quarterly*, *49*(2), 139–153. https://doi.org/10.1177/001698620504900204

Gagné, F. (1985). Giftedness and Talent: Reexamining a Reexamination of Definitions. *Gifted Child Quarterly, 29*(3), 103–112.

Galton, F. (1892). Hereditary genius. London: Macmillan and Co.

*Geddes, K. A. (2011). Academic dishonesty among gifted and high-achieving students. *Gifted Child Today, 34*(2), 50–56. https://doi.org/10.1177/107621751103400214

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 213–228. https://doi.org/10.1080/02671520701296189

*Griffin, K. A., Allen, W. R., Kimura-Walsh, E. & Yamamura, E. K. (2007). Those who left, those who stayed: Exploring the educational opportunities of high-achieving black and latina/o students at magnet and nonmagnet Los Angeles high schools (2001-2002). *Educational Studies: Journal of the American Educational Studies Association, 42*(3), 229–247.

Hany, E. A. (2012). Zum Verhältnis von Begabung und Leistung. In A. Hackl, C. Pauly, O. Steenbuck & G. Weigand (Hrsg.), *Werte schulischer Begabtenförderung. Begabung und Leistung (Karg-Hefte. Beiträge zur Begabtenförderung und Begabungsforschung*), 35–40. Frankfurt, M.: Karg-Stiftung. https://doi.org/10.25656/01:9030*Harpalani, V. (2017). Counterstereotypic identity among high-achieving black students. *Penn GSE Perspectives on Urban Education, 14*(1), 1–9.

Heckhausen, H. (1974). *Leistung und Chancengleichheit* (Motivationsforschung, Bd. 2). Göttingen: Hogrefe.

Heller, K. A. (Hrsg.). (2001). *Hochbegabung im Kindes- und Jugendalter* (2. Aufl.). Göttingen: Hogrefe.

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

Hoover, H. D., Dunbar, S. B. & Frisbie, D. A. (2001). *Iowa tests of basic skills*. Itasca, IL: Riverside Publishing.

*Huang, H. & Zhu, H. (2017). High achievers from low socioeconomic backgrounds: The critical role of disciplinary climate and grit. *Mid-Western Educational Researcher*, *29*(2), 93–116.

Hussmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M. et al. (Hrsg.). (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

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

*Kaufman, S. B. (2009). Faith in intuition is associated with decreased latent inhibition in a sample of high-achieving adolescents. Psychology of Aesthetics, Creativity, and the Arts, 3(1), 28–34. https://doi.org/10.1037/a0014822

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2007). *Zur Entwicklung nationaler Bildungsstandards. Expertise* (Bildungsforschung Band 1). Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Köller, O. & Baumert, J. (2017). Hochleistende Schülerinnen und Schüler im mehr-und zweigliedrigen System. In M. Neumann, M. Becker, J. Baumert, K. Maaz & O. Köller (Hrsg.), *Zweigliedrigkeit im deutschen Schulsystem. Potenziale und Herausforderungen in Berlin* (1. Aufl., S. 227). Münster: Waxmann Verlag GmbH.

*Kour, S. (2015). Scientific temper among academically high and low achieving adolescent girls. *Journal of Education and Practice*, *6*(34), 96–101.

Krajewski, K., Dix, S. & Schneider, W. (2020). *Deutscher Mathematiktest für zweite Klassen* (2. Aufl.). Hogrefe.

*Lauen, D. L. & Gaddis, S. M. (2016). Accountability pressure, academic standards, and educational triage. *Educational Evaluation and Policy Analysis*, *38*(1), 127–147. https://doi.org/10.3102/0162373715598577

Lenhard, W., Lenhard, A. & Schneider, W. (2020). *Ein Leseverständnistest für Erst- bis Siebtklässler – Version II* (4. Aufl.). Göttingen: Hogrefe.

Lintorf, K. (2012a). Messtheoretische Güte von Schulnoten. In K. Lintorf (Hrsg.), *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale* (S. 37–66). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94339-8_3

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

Lubinski, D. (2016). From Terman to today. *Review of Educational Research*, *86*(4), 900–944. https://doi.org/10.3102/0034654316675476

Lubinski, D. & Benbow, C. P. (2006). Study of Mathematically Precocious Youth After 35 Years: Uncovering Antecedents for the Development of Math-Science Expertise. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, *1*(4), 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

*Lüftenegger, M., Kollmayer, M., Bergsmann, E., Jöstl, G., Spiel, C. & Schober, B. (2015). Mathematically gifted students and high achievement: The role of motivation and classroom structure. *High Ability Studies*, *26*(2), 227–243. https://doi.org/10.1080/13598139.2015.1095075

*Ma, X. (2005). A longitudinal assessment of early acceleration of students in mathematics on growth in mathematics achievement. *Developmental Review, 25*(1), 104–131. https://doi.org/10.1016/j.dr.2004.08.010

Maaz, K., Baeriswyl, F. & Trautwein, U. (2013). „Herkunft zensiert?“ Leistungsdiagnostik und soziale Ungleichheiten in der Schule [“Origin graded?“ Performance Diagnostics and Social Inequalities at School]. In D. Deißner (Hrsg.), *Chancen bilden. Wege zu einer gerechteren Bildung - ein internationaler Erfahrungsaustausch* (S. 185–188). Wiesbaden: Springer Fachmedien Wiesbaden.

*Marsh, K. (2013). "Staying Black": The demonstration of racial identity and womanhood among a group of young high-achieving black women. *International Journal of*

Qualitative Studies in Education (QSE), 26(10), 1213–1237. https://doi.org/10.1080/09518398.2012.731536

McBee, M. T. (2006). A descriptive analysis of referral sources for gifted identification screening by race and socioeconomic status. *The Journal of Secondary Gifted Education*, *XVII*(2), 103–111.

*McGee, E. O. & Pearman, F. Alvin, II. (2015). Understanding black male mathematics high achievers from the inside out: Internal risk and protective factors in high school. *Urban Review: Issues and Ideas in Public Education, 47*(3), 513–540. https://doi.org/10.1007/s11256-014-0317-2

*Mioduser, D. & Betzer, N. (2007). The contribution of project-based-learning to high-achievers‘ acquisition of technological knowledge and skills. *International Journal of Technology and Design Education*, *18*(1), 59–77. https://doi.org/10.1007/s10798-006-9010-4

Möller, J. & Pohlmann, B. (2010). Achievement differences and self-concept differences. Stronger associations for above or below average students? *The British Journal of Educational Psychology*, *80*(Pt 3), 435–450. https://doi.org/10.1348/000709909X485234

Moser Opitz, E., Stöckli, M., Grob, U., Nührenbörger, M. & Reusser, L. (2019). BASIS-MATH-G 3+ Gruppentest zur Basisdiagnostik Mathematik für das vierte Quartal der 3. Klasse und das erste Quartal der 4. Klasse. Hogrefe.

*Mourgues, C. V., Hein, S., Tan, M., Diffley III, R. & Grigorenko, E. L. (2016). The role of noncognitive factors in predicting academic trajectories of high school students in a selective private school. *European Journal of Psychological Assessment*, *32*(1), 84–94. https://doi.org/10.1027/1015-5759/a000332

*Neuendorf, C., Jansen, M. & Kuhl, P. (2020). Competence development of high achievers within the highest track in German secondary school: Evidence for Matthew effects or compensation? *Learning and Individual Differences*, *77*, 101816. https://doi.org/10.1016/j.lindif.2019.101816

Neuendorf, C., Kuhl, P. & Jansen, M. (2017). Leistungsstarke Schülerinnen und Schüler in Deutschland. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich.* Münster: Waxmann.

*Obergriesser, S. & Stoeger, H. (2016). The influence of emotions and learning preferences on learning strategy use before transition into high-achiever track secondary school. *High Ability Studies*, *27*(1), 5–38. https://doi.org/10.1080/13598139.2015.1100980

OECD. (2016). *PISA 2015 Ergebnisse (Band 1). Exzellenz und Chancengerechtigkeit in der Bildung*. Bielefeld: PISA, W. Bertelsmann Verlag. https://doi.org/10.3278/6004573w

OECD. (2015). *The ABC of Gender Equality in Education. Aptitude, behaviour, confidence*. OECD Publishing. https://doi.org/10.1787/9789264229945-en

OECD. (2009). *Top of the class. High performers in science in PISA 2006*. OECD Publishing. https://doi.org/10.1787/9789264060777-en

Pant, H. A., Tiffin-Richards, S. P. & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment. Projekt Standardsetting. *Zeitschrift für Pädagogik*, *56*(Beiheft), 175–188.

*Parsons, E. (2016). Does attending a low-achieving school affect high-performing student outcomes? *Teachers College Record, 118*(8).

*Peters, M. P. & Bain, S. K. (2011). Bullying and victimization rates among gifted and high-achieving students. *Journal for the Education of the Gifted, 34*(4), 624–643. https://doi.org/10.1177/016235321103400405

Preckel, F., Golle, J., Grabner, R., Jarvin, L., Kozbelt, A., Müllensiefen, D., … Worrell, F. C. (2020). Talent development in achievement domains: A psychological framework for within- and cross-domain research. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 691–722. https://doi.org/10.1177/1745691619895030

Prenzel, M., Schütte, K. & Walter, O. (2007). Interesse an den Naturwissenschaften. In M. Prenzel, C. Artelt, J. Baumert, W. Blum & M. Hammann (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 107–124). Münster: Waxmann.

*Qin, D. B., Rak, E., Rana, M. & Donnellan, M. B. (2012). Parent-child relations and psychological adjustment among high-achieving Chinese and European American adolescents. *Journal of Adolescence, 35*(4), 863–873. https://doi.org/10.1016/j.adolescence.2011.12.004

*Rambo-Hernandez, K. E. & McCoach, D. B. (2015). High-achieving and average students' reading growth. Contrasting school and summer trajectories. *The Journal of Educational Research*, *108*(2), 112–129. https://doi.org/10.1080/00220671.2013.850398

*Rathod, A. (2010). Self-regulated learning of high achievers. *Journal on Educational Psychology*, *4*(2), 33–38.

*Reilly, D., Neumann, D. L. & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, *107*(3), 645–662. https://doi.org/10.1037/edu0000012

Reiss, K. & Sälzer, C. (2016). Fünfzehn Jahre Pisa: Bilanz und Ausblick. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015. Eine Studie zwischen Kontinuität und Innovation* (S. 375–381). Münster: Waxmann.

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E. & Köller, O. (Hrsg.). (2016). *PISA 2015. Eine Studie zwischen Kontinuität und Innovation*. Münster: Waxmann. Verfügbar unter: http://www.content-select.com/index.php?id=bib_view&ean=9783830985556

Reiss, K., Weis, M. & Klieme, E. (2019). PISA 2018. Grundbildung im internationalen Vergleich.

*Robinson, N. M., Lanzi, R. G., Weinberg, R. A., Ramey, S. L. & Ramey, C. T. (2002). Family factors associated with high academic competence in former Head Start children at third grade. *Gifted Child Quarterly*, *46*(4), 278–290. https://doi.org/10.1177/001698620204600404

*Roos, A.-L., Bieg, M., Goetz, T., Frenzel, A. C., Taxer, J. & Zeidner, M. (2015). Experiencing more mathematics anxiety than expected? Contrasting trait and state anxiety in high achieving students. *High Ability Studies*, *26*(2), 245–258. https://doi.org/10.1080/13598139.2015.1095078

Rost, D. H. (Hrsg.). (2009). *Hochbegabte und hochleistende Jugendliche. Befunde aus dem Marburger Hochbegabtenprojekt* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 72, 2., erw. Aufl.). Münster: Waxmann.

Rost, D. H. & Buch, S. (2018). Hochbegabung. In D. H. Rost, J. R. Sparfeldt & S. Buch (Hrsg.), *Handwörterbuch pädagogische Psychologie* (Beltz Psychologie 2018, 5., überarbeitete und erweiterte Auflage, S. 226–242). Weinheim: Beltz.

Rothenbusch, S., Zettler, I., Voss, T., Lösch, T. & Trautwein, U. (2016). Exploring reference group effects on teachers' nominations of gifted students. *Journal of Educational Psychology*, *108*(6), 883–897. https://doi.org/10.1037/edu0000085

Rüdiger, C., Jansen, M. & Rjosk, C. (2021). Empirische Arbeit: „Paul ist nicht so gut in Deutsch". Geschlechtsdifferenzielle Benotung im Fach Deutsch – eine Sekundäranalyse der Daten des IQB-Bildungstrends 2015. *Psychologie in Erziehung und Unterricht*, *68*. https://doi.org/10.2378/peu2021.art08d

*Rutkowski, D., Rutkowski, L. & Plucker, J. A. (2012). Trends in education excellence gaps: A 12-year international perspective via the multilevel model for change. *High Ability Studies, 23*(2), 143–166. https://doi.org/10.1080/13598139.2012.735414

*Salmela, M. & Uusiautti, S. (2015). A positive psychological viewpoint for success at school – 10 characteristic strengths of the Finnish high-achieving students. *High Ability Studies*, *26*(1), 117–137. https://doi.org/10.1080/13598139.2015.1019607

Schrader F. W., Helmke A. (2008) Determinanten der Schulleistung. In: Schweer M.K.W. (eds) Lehrer-Schüler-Interaktion. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91104-5_11

Schurtz, I. M., Pfost, M. & Artelt, C. (2014). Variieren die Selbstkonzeptdifferenzen in Abhängigkeit vom Leistungsniveau? Differenzielle Zusammenhänge in Deutsch, Englisch und Mathematik. *Zeitschrift für Pädagogische Psychologie*, *28*(1-2), 31–42. https://doi.org/10.1024/1010-0652/a000122

Schwippert, K., Kasper, D., Köller, O., McElvany, N., Selter, C., Steffensky, M. et al. (Hrsg.). (2020). TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich (1. Auflage). Münster: Waxmann.

Simons, D. J., Shoda, Y. & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed addition to all empirical papers. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

*Sontag, C. & Stoeger, H. (2015). Can highly intelligent and high-achieving students benefit from training in self-regulated learning in a regular classroom context? *Learning and Individual Differences*, *41*, 43–53. https://doi.org/10.1016/j.lindif.2015.07.008

*Sparfeldt, J. R., Buch, S. R. & Rost, D. H. (2010). Klassenprimus bei durchschnittlicher Intelligenz. *Zeitschrift für Pädagogische Psychologie*, *24*(2), 147–155. https://doi.org/10.1024/1010-0652/a000012

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S. & Henschel, S. (Hrsg.). (2019). Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich. Münster: Waxmann.

Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

*Steinmayr, R. & Spinath, B. (2017). Why time constraints increase the gender gap in measured numerical intelligence in academically high achieving samples. European Journal of Psychological Assessment. https://doi.org/10.1027/1015-5759/a000400

*Stoeger, H., Hopp, M. & Ziegler, A. (2017). Online mentoring as an extracurricular measure to encourage talented girls in STEM (science, technology, engineering, and mathematics): An empirical study of one-on-one versus group mentoring. Gifted Child Quarterly, 61(3), 239–249. https://doi.org/10.1177/0016986217702215

Subotnik, R. F., Olszewski-Kubilius, P. & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *12*(1), 3–54. https://doi.org/10.1177/1529100611418056

Terman, L. M. (1954). The discovery and encouragement of exceptional talent. *American Psychologist*, *9*, 221–230. https://doi.org/10.1037/h0060516

Terman, L. M. (1926). Genetic studies of genius. Mental and physical traits of a thousand gifted children. Stanford University Press.

Trautwein, U., Köller, O. & Kämmerer, E. (2002). Effekte innerer und äußerer Leistungsdifferenzierung auf selbstbezogene Fähigkeitskognitionen, die wahrgenommene Unterrichtspartizipation und die wahrgenommene soziale Akzeptanz. *Psychologie in Erziehung und Unterricht*, *49*(4), 273–286.

*Vock, M., Köller, O. & Nagy, G. (2013). Vocational interests of intellectually gifted and highly achieving young adults. *The Journal of Educational Psychology*, *83*, 305–328. https://doi.org/10.1111/j.2044-8279.2011.02063.x

*Walker, C. L. & Shore, B. M. (2015). Myth busting: Do high-performance students prefer working alone? *Gifted and Talented International*, *30*(1-2), 85–105. https://doi.org/10.1080/15332276.2015.1137461

*Wang, M.-T., Eccles, J. S. & Kenny, S. (2013). Not lack of ability but more choice: individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science*, *24*(5), 770–775. https://doi.org/10.1177/0956797612458937

*Weckbacher, L. M. & Okamoto, Y. (2012). Spatial experiences of high academic achievers: insights from a developmental perspective. *Journal for the Education of the Gifted, 35*(1), 48–65. https://doi.org/10.1177/0162353211432038

Woodcock, R. W., McGrew, K. S. & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.

*Yong, B. C. S. (2012). Comparison between the thinking styles of students in a science school and a mainstream school. *Journal of Science and Mathematics Education in Southeast Asia, 35*(1), 60–83.

*Zhou, Y., Fan, X., Wei, X. & Tai, R. H. (2017). Gender gap among high achievers in math and implications for STEM pipeline. *Asia-Pacific Education Researcher*, *26*(5), 259–269. https://doi.org/10.1007/s40299-017-0346-1

Ziegler, A. & Raul, T. (2000). Myth and Reality: A review of empirical studies on giftedness. *High Ability Studies*, *11*(2), 113–136. https://doi.org/10.1080/13598130020001188

Zimmer, K., Brunner, M., Lüdtke, O., Prenzel, M. & Baumert, J. (2007). Die PISA-Spitzengruppe in Deutschland: Eine Charakterisierung hochkompetenter Jugendlicher. In K. A. Heller & A. Ziegler (Hrsg.), *Begabt sein in Deutschland* (Talentförderung, Expertiseentwicklung, Leistungsexzellenz, Bd. 1, S. 193–208). Berlin: Lit.

Anhang I-A

Tabelle I-A1

Suchstrategie bei der Literaturrecherche in Fachdatenbanken

Datenbank	Hauptsuchbegriffe	Filter	*N*
PsychARTICLES	„high-achiev*“		55
	“high-achiev*” OR “high achiev”* AND students		4
Psyndex Literature & AV Media	leistungsstark*	TYPE: Academic Journals	52
ERIC	high achievers AND students	Filter: - wissenschaftliche Zeitschriften - high achievement (Thema) - NOT(Early Childhood Education, Higher Education, Kindergarten, Two Year Colleges) (Bildungsetappen)	146
	high performing students	Filter: - wissenschaftliche Zeitschriften - high achievement (Thema) NOT(Early Childhood Education, Higher Education, Kindergarten, Two Year Colleges) (Bildungsetappen)	28

Tabelle I-A2

Kurzüberblick über eingeschlossene Studien

Autor, Jahr	N	Altersgruppe	Forschungsdesign	Land	Indikatoren	Fachbezug	Bezugsnorm
Adelodun et al., 2015	50	O	quantitativ	Nigeria	Noten	Einzelfach	-
Allen & Griffin, 2006	17	M - O	qualitativ	USA	Noten, Kontext	Gesamtwert	kriterial
Assouline et al., 2017	1146	M	experimentell	USA	Test	Gesamtwert	sozial (P)
Bütüner & Filiz, 2017	233	G	qualitativ	Türkei	Noten	Einzelfach	kriterial
Barron, 2000	96	M	experimentell	USA	Schulform	-	kriterial
Bergold et al., 2017	74868	G	quantitativ	Multinational (EU)	LSA	Kombination	kriterial
Bergold et al., 2020	3928	G	quantitativ	Deutschland	LSA	Kombination	sozial (P)
Berkowitz & Cicchelli, 2004	63	M	quantitativ	USA	Noten	Einzelfach	kriterial
Carroll & Bailey, 2016	967	G	quantitativ	USA	LSA	Kombination	kriterial
Carter Andrews, 2012	2181	M - O	qualitativ	USA	Noten, Einschätzung, Kontext, Andere	Gesamtwert	sozial (S)
Carter, 2008	9	O	qualitativ	USA	Noten, Kontext, Andere	Gesamtwert	kriterial
Castejón & Zanjaco, 2015	515655	M	quantitativ	Multinational	LSA	Einzelfach	sozial (P)
Cheung, 2017	22636	M	quantitativ	Multinational (Asien)	LSA	Einzelfach	sozial (P)
Clausen et al., 2013	53	M	qualitativ	Deutschland	Schulform	-	kriterial
Crawford et al., 2017	480653	G	quantitativ	Großbritannien	LSA	Einzelfach	kriterial
Ee et al., 2003	566	G	quantitativ	Singapur	Schulform	-	sozial (P)
Erbas & Bas, 2015	217	M	quantitativ	Türkei	Schulform	Kombination	kriterial
Flores-Gonzalez, 2005		O	qualitativ	USA	Noten, Kontext	Gesamtwert	kriterial
Forgasz & Hill, 2013	4811	O	qualitativ	Australien	Noten	Einzelfach	sozial (P)

Autor, Jahr	N	Altersgruppe	Forschungsdesign	Land	Indikatoren	Fachbezug	Bezugsnorm
Gagné, 2016 (2)		G - M	quantitativ	USA	Test	Gesamtwert	sozial (P)
Geddes, 2011		M - O	quantitativ	USA	Kursstufe	Kombination	kriterial
Griffin et al., 2007	34	O	qualitativ	USA	Noten, Kontext	Gesamtwert	kriterial
Harpalani, 2017	779	M	quantitativ	USA	Noten	Gesamtwert	kriterial
Huang & Zhu, 2017 (2)	1220	M	quantitativ	USA	LSA	Einzelfach	sozial (P)
Kaufman, 2009	162	O	experimentell	Großbritannien	Schulform	Gesamtwert	kriterial
Kour, 2015	120	O	quantitativ	Indien	Noten	-	kriterial
Lauen & Gaddis, 2016 (2)	500000	G - M	quantitativ	USA	LSA	Einzelfach	sozial (P)
Lüftenegger et al., 2015	210	M - O	quantitativ	Österreich	Noten	Einzelfach	kriterial
Ma, 2005	3116	M	quantitativ	USA	Test	Einzelfach	sozial (S)
Marsh, 2013	9	O	qualitativ	USA	Schulform	Kombination	kriterial
McGee & Pearman, 2015	13	M - O	qualitativ	USA	Kursstufe	Einzelfach	sozial (K)
Mioduser & Betzner, 2008	120	O	experimentell	Israel	Noten	-	-
Mourgues et al., 2016	8586	M	quantitativ	USA	Schulform, Noten	Gesamtwert	sozial (S)
Neuendorf et al., 2020 (2)	1010	M	quantitativ	Deutschland	Test	Einzelfach	sozial (S)
Obergriesser & Stoeger, 2016	200	G	quantitativ	Deutschland	Noten	Mittelwert	kriterial
Parsons, 2016		G	quantitativ	USA	LSA	Einzelfach	sozial (P)
Peters & Bain, 2011	90	M - O	quantitativ	USA	Kursstufe	-	sozial (K)
Qin et al., 2012	487	M	quantitativ	USA	Schulform	-	kriterial
Rambo-Hernandez & McCoach, 2015 (1)	21021	G	quantitativ	USA	Test	Einzelfach	sozial (K)
Rambo-Hernandez & McCoach, 2015 (2)	70521	G	quantitativ	USA	Test	Einzelfach	sozial (P)
Rathod, 2010	480	O	quantitativ	Indien	Noten	-	kriterial
Reilly et al., 2015 (2)		G - O	quantitativ	USA	LSA	Einzelfach	kriterial

Autor, Jahr	N	Altersgruppe	Forschungsdesign	Land	Indikatoren	Fachbezug	Bezugsnorm
Robinson et al., 2002	5400	G	mixed methods	USA	Test	Kombination	sozial (S)
Roos et al., 2015	237	M	quantitativ	Deutschland	Noten	Einzelfach	kriterial
Rutkowski et al., 2012 (2)	272000	M	quantitativ	Multinational	LSA	Einzelfach	kriterial
Salmela & Uusiautti, 2015	14	O	qualitativ	Finnland	Noten	Gesamtwert	kriterial
Sontag & Stoeger, 2015	123	G	experimentell	Deutschland	Noten	Mittelwert	sozial (S)
Sparfeldt et al. 2010	256	M	quantitativ	Deutschland	Noten, Lehrkräfte-einschätzungen	Mittelwert	kriterial
Steinmayr & Spinath, 2017	666	O	quantitativ	Deutschland	Schulform	-	kriterial
Stoeger et al., 2017	347	M - O	experimentell	Deutschland	Schulform	-	kriterial
Vock et al., 2013	4694	O	quantitativ	Deutschland	Noten	Gesamtwert	sozial (S)
Walker & Shore, 2015	69	G	quantitativ	Kanada	Schulform	-	kriterial
Wang et al., 2013	1490	O	quantitativ	USA	Test	Kombination	sozial (S)
Weckbacher & Okamoto, 2012	43	M	quantitativ	USA	Noten, Test, Kontext	Kombination	sozial (P)
Yong, 2012	378	M	quantitativ	Brunei	Schulform	Einzelfach	kriterial
Zhou et al., 2017 (4)	130400	M	quantitativ	Multinational	LSA	Einzelfach	sozial (P)

Anmerkungen. Eine vollständige Übersicht über die eingeschlossenen Studien sowie die Kodierung aller im Artikel verwendeter Merkmale wird unter osf.io/jzkv6/ bereitgestellt.
Altersgruppe: O = Oberstufe/Sekundarstufe II, M=Mittelstufe/Sekundarstufe I, G=Grundschule. LSA = Large Scale Assessment. P = Populationsebene, S=Stichprobenebene, K = Klassenebene.

Tabelle I-A3

In den Studien verwendete Notensysteme und Noten-Cut-offs

Land	Notensystem	Cut-off-Werte
USA	GPA	2.8[a]; 3[b]; 4[c]
	0-100 %	95 %[c]
	Letter scale F-A	B[d]; C[e]
Australien	0-50 Punkte	46[f]
Indien	0-100 %	60 %[g]; 70%[h]
Türkei	1-5	4[i]
Österreich	5-1	1[j]
Deutschland	6-1	1.4[k]; 1.75[l]; 2.33[m]
Finnland	Letter scale I-L	L[n]

Anmerkung. Dargestellt sind die in den Studien genutzten Notenwerte und ihre Cut-offs. Die Werte basieren auf 17 Studien.

[a]Carter Andrews, 2012; Carter, 2008. [b]Allen & Griffin, 2006; Griffin et al., 2007. [c]Weckbacher & Okamoto, 2012; Berkowitz & Chicchelli, 2004. [d] Harpalani, 2017. [e]Flores-Gonzales, 2005. [f]Forgasz & Hill, 2013. [g]Rahtod, 2010. [h]Kour, 2015. [i]Bütüner & Filiz, 2017. [j]Lüftenegger, Kollmayer, Bergsmann, Jöstl, Spiel & Schober, 2015. [k]Sparfeldt, Buch & Rost, 2010. [l]Roos, Bieg, Goetz, Frenzel, Taxer & Zeidner, 2015. [m]Obergriesser & Stoeger, 2016. [n]Salmela & Uusiautti, 2015.

1.5. Zusammenfassung und Zwischenfazit aus Beitrag I

1.5.1. Zusammenfassung der Ergebnisse

Die Untersuchung von Operationalisierungen von Leistungsstärke in 55 wissenschaftlichen Artikeln der Jahre 2000 – 2020 zeigte: 1) Die meistgenutzten Indikatoren für Leistungsstärke waren Tests und Noten, wobei in der Mehrzahl der Fälle lediglich ein einzelner Indikator genutzt wurde. 2) Entgegen der Annahmen, die man auf Basis der in Kapitel 1 vorgestellten Theorien treffen könnte (nämlich, dass Leistung im Gegensatz zur Begabung stärker fachbezogen untersucht wird), waren sowohl fachspezifische als auch fächerübergreifende Konzepte von Leistungsstärke beide häufig vertreten. 3) Das Signifikanzkriterium variierte stark, allerdings wurden in der Hälfte der Studien weniger als 10 % der Population als leistungsstark eingeschätzt. 4) Zusatzauswertungen zeigten, dass die Operationalisierung stark abhängig von dem methodischen Zugang der Studien war. Qualitative Untersuchungen boten ganz andere Möglichkeiten der Operationalisierung (Vorhandensein einer großen Fülle an individuellen und kontextspezifischen Informationen, multikriteriale und intersubjektive Betrachtungsweise) als Large-Scale-Assessment-Studien (mit objektiven, standardisierten, repräsentativen Daten und großen Stichprobenumfängen) oder (quasi-)experimentelle Untersuchungen (bei denen das Kriterium passgenau für die jeweilige Fragestellung erhoben werden kann). Gleichzeitig waren einige der inhaltlichen Fragestellungen, die von den Studien untersucht wurden, klar einem bestimmten methodischen Zugang zugeordnet. So nutzten beispielsweise Studien zu Disparitäten eher Large-Scale-Assessmentdaten; leistungsstarke Angehörige ethnischer Minderheiten wurden überwiegend mit einem qualitativen Ansatz untersucht; Studien zu Lernstrategien oder Unterrichtsmethoden nutzten eher kleinere quantitative z.T. (quasi-)experimentelle Untersuchungsdesigns.

Es zeigte sich auch, dass ein Teil der Studien Operationalisierungen innerhalb der Untersuchung entweder als Robustheits- oder als Generalisierbarkeitsüberprüfung variierte und unterschiedliche Bezugsnormen, Signifikanzkriterien oder Fächerdomänen verglich. Auf dieser Beobachtung aufbauend wird hier das Potenzial von Large-Scale-Assessments hervorgehoben, durch Nutzung unterschiedlicher Operationalisierungen einerseits die Generalisierbarkeit von Befunden zu prüfen und andererseits Phänomene mit Bezug zu Leistungsstärke zu explorieren.

1.5.2. Folgerungen für das weitere Vorgehen

Aus den Ergebnissen des ersten Beitrags folgt, dass Leistungsstärke bisher in der Forschung sehr heterogen verstanden und operationalisiert wird. Auch wenn es bei manchen Fragestellungen eine gute Begründung zur Wahl eines bestimmten Indikators oder einer Referenznorm gibt, sind doch

einige, wenn nicht gar die meisten Aspekte der Operationalisierung häufig nicht eindeutig aus der Theorie ableitbar. Innerhalb des Rahmens, der durch Fragestellung und Methodik einer Untersuchung gegeben ist, wäre also oftmals eine Vielzahl an gleichberechtigten Operationalisierungen denkbar. Diese Situation kann dazu führen, dass Fragen der Generalisierbarkeit und der Anschlussfähigkeit unterschiedlicher Studien aufgeworfen werden. Forschende sehen sich daher mit der Anforderung konfrontiert, im Forschungsprozess eine Vielzahl von Entscheidungen treffen zu müssen, die keine eindeutig beste Lösung haben. Dies wird auch mit dem Ausdruck „Garden of Forking Paths“ (Gelman & Loken, 2013; Gelman & Loken, 2014) beschrieben. Die damit verbundene Gefahr, dass Ergebnisse sich nur bei einer ganz bestimmten Spezifikation („knife-edge specification“, Young & Holsteen, 2017) zeigen, wird im Rahmen der Replikationskrise in der Psychologie als ein Grund dafür diskutiert, dass Ergebnisse von Studien sich teilweise nicht leicht replizieren lassen. Dies gilt insbesondere, wenn Entscheidungen während des Analyseprozesses nicht transparent und nachvollziehbar dokumentiert werden. Als Möglichkeit, mit diesen Freiheitsgraden umzugehen wurde von verschiedenen Forschenden ein Vorgehen vorgeschlagen, welches unter dem Begriff „Multiversumsanalyse“ zusammengefasst werden kann (Rohrer, 2021; Simonsohn, Simmons & Nelson, 2015; Steegen, Tuerlinckx, Gelman & Vanpaemel, 2016). Dabei wird zunächst für den Datenaufbereitungs- und Analyseprozess eine Anzahl von Prozessschritten identifiziert, bei denen alternative Operationalisierungen möglich und verteidigbar wären. Für jeden dieser Schritte werden unterschiedliche mögliche Vorgehensweisen festgelegt. In der Analyse wird dann jeder mögliche Pfad durch diesen Entscheidungsbaum gegangen, d.h. es wird ein Multiversum an aufbereiteten Datensätzen (Steegen et al., 2016) beziehungsweise ein Multiversum an Resultaten generiert. Ziel der Multiversumsanalyse ist es, solche Spezifikationen zu identifizieren, welche die Resultate in besonderem Maße beeinflussen, sei es, a) um einen zuverlässigeren Schluss von den Analysen auf die Forschungsfragen zu treffen (Simonsohn et al., 2015) oder b) potenziell theoretisch gehaltvolle Aspekte der Operationalisierung zu identifizieren (Steegen et al., 2016). Diese beiden Ansprüche an die Multiversumsanalyse zeigen auch, dass sie sowohl im Entdeckungszusammenhang, also mit dem Ziel der Exploration als auch im Begründungszusammenhang, mit dem Ziel einer validen Inferenz eingesetzt werden kann.

Bei Jansen, Neuendorf und Kocaj (2021) wird diskutiert, dass dieser Ansatz innerhalb des Programms des kritischen Multiplismus (Patry, 2013; Shadish, 1986, 1993) verortet werden kann. „*Der kritische Multiplismus als Forschungsprogramm plädiert für den Einsatz einer Vielzahl an Operationalisierungen, Analysemethoden, Studien, Hypothesen und Theorien* [zur Untersuchung einer Fragestellung]. *Die Multiversumsanalyse ordnet sich in diesen historischen Ansatz ein,*

indem multiple Operationalisierungen und Auswertungsstrategien angewandt werden" (Jansen et al., 2021). Die Autoren halten auch fest, dass sich für dieses Vorgehen Large-Scale-Assessmentdaten in besonderem Maße eignen, da eine Vielzahl an Messpunkten (Variablen und Personen) enthalten sind, die die Spezifikation einer größeren Zahl an Alternativmodellen erlaubt.

In der vorliegenden Arbeit wird die Methode der Multiversumsanalyse auf zwei Fragestellungen mit Bezug zu Leistungsstärke angewendet. Während es bei der ersten Fragestellung um die differenzielle Leistungsentwicklung leistungsstarker Schülerinnen und Schüler am Gymnasium geht, dreht sich die zweite Fragestellung um deren soziale Integration. Der erste der beiden inhaltlichen Beiträge berührt also die Frage ob das Bildungssystem die Lernpotenziale der Leistungsspitze derart unterstützt, dass ein kumulativer Vorteil für leistungsstärkere Kinder und Jugendliche entsteht. Die zweite Frage ist eher im Mikrokontext verortet. Hier geht es um die soziale Einbindung Leistungsstarker in ihren Klassenkontext. Obwohl es qualitative Berichte und Stereotypenforschung gibt, die (zum Teil auch genderspezifische) Stigmatisierung von Leistungsexzellenz nahelegen (Gronostaj, Werner, Bochow & Vock, 2016; Pelkner, Günther & Boehnke, 2002; Pelkner & Boehnke, 2003), zeigt sich in Studien, welche sich direkt mit Freundschaftsnetzwerken in Schulklassen beschäftigen, meist ein positiver Effekt der Leistung auf die tatsächliche soziale Eingebundenheit (Wentzel, Jablansky & Scalise, 2021). Im Beitrag wird der Frage nachgegangen, inwiefern unterschiedliche Operationalisierungen von Leistungsstärke und sozialer Integration diese widersprüchlichen Befunde erklären können. Bei beiden Fragestellungen soll die Multiversumsanalyse als Instrument eingesetzt werden, um, wie in Artikel I vorgeschlagen, die Operationalisierung von Leistungsstärke zu variieren und Effekte auf die Ergebnisse der Studien zu untersuchen. Es werden beide Akzentuierungen der Multiversumsanalyse verwendet, das heißt, sie wird sowohl im Begründungszusammenhang im Rahmen konfirmatorischer Analysen eingesetzt als auch im Entdeckungszusammenhang als Explorationswerkzeug. In Studie II zum kumulativen Vorteil Leistungsstarker wird im Rahmen der Multiversumsanalyse unter anderem das Signifikanzkriterium für Leistungsstärke verschoben, d.h. es werden unterschiedliche Anteile der Stichprobe als leistungsstark klassifiziert. Dies soll eine Beurteilung erlauben, wie stark die Studienergebnisse von dieser zu einem gewissen Grad willkürlichen Entscheidung abhängen. Neben der Größe der Gruppe der Leistungsstarken werden auch weitere Aspekte des Analysemodells variiert. Damit lässt sich der Effekt der Leistungsgruppierung in Relation setzen zum Effekt anderer Analyseentscheidungen. Der Fokus der Multiversumsanalyse besteht also bei diesem Beitrag stärker darin, die Replizierbarkeit des Ergebnisses der Hauptanalyse bei unterschiedlichen Spezifikationen zu untersuchen. In Artikel III hingegen wird das Potenzial der Multiversumsanalyse im Entdeckungszusammenhang

demonstriert. Hierbei werden nicht austauschbare Entscheidungen nebeneinandergestellt - vielmehr werden bewusst Operationalisierungen kontrastiert um die übergeordnete Fragestellung, die soziale Integration Leistungsstarker, in ihren unterschiedlichen Facetten zu beleuchten und Muster zu identifizieren, die einer näheren Untersuchung würdig sind und Grundlage einer möglichen späteren Theoriebildung sein könnten. Beiden Artikeln ist gemeinsam, dass keine Inferenz aus den Ergebnissen der Multiversumsanalyse betrieben wird (wie von Simonsohn et al., 2015 vorgeschlagen), sondern sie eher als Werkzeug eingesetzt wird, um mögliche Varianzquellen von Effekten aufzuspüren.

Literaturverzeichnis

Bundesministerium für Bildung und Forschung; Kultusministerkonferenz. (2016, 28. September). *Gemeinsame Initiative von Bund und Ländern zur Förderung leistungsstarker und potenziell besonders leistungsfähiger Schülerinnen und Schüler*.

Eickelmann, B., Bos, W., Gerick, J. & Labusch, A. (2019). Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern der 8. Jahrgangsstufe in Deutschland im zweiten internationalen Vergleich. In B. Eickelmann, W. Bos, J. Gerick, F. Goldhammer, H. Schaumburg, K. Schwippert et al. (Hrsg.), *ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking* (S. 113–136). Münster: Waxmann.

Eurydice-Netz. (2009). *Nationale Lernstandserhebungen von Schülern in Europa. Ziele, Aufbau und Verwendung der Ergebnisse*. Luxemburg, Brüssel: Exekutivagentur Bildung, Audiovisuelles und Kultur.

Forschungsverbund LemaS. (2018). *Welcher Leistungsbegriff liegt "Leistung macht Schule" zugrunde?* Verfügbar unter: https://www.leistung-macht-schule.de/files/LemaS_Leistungsbegriff.pdf

Gelman, A. & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem*: Columbia University. Verfügbar unter: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A. & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, *102*(6), 460–465. Verfügbar unter: http://www.jstor.org/stable/43707868

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

Gronostaj, A., Werner, E., Bochow, E. & Vock, M. (2016). How to Learn Things at School You Don't Already Know. *Gifted Child Quarterly*, *60*(1), 31–46. https://doi.org/10.1177/0016986215609999

Ingenkamp, K. (1977). Sind Zensuren aus verschiedenen Klassen vergleichbar? In K. Ingenkamp (Hrsg.), *Die Fragwürdigkeit der Zensurengebung* (7. Aufl., S. 194–201). Weinheim: Beltz.

Jansen, M., Neuendorf, C. & Kocaj, A. (2021). Welche Potenziale bieten Sekundäranalysen für die Erhöhung von Forschungsqualität und Replizierbarkeit. *Zeitschrift für Pädagogik*, *67*(6), 840–859. https://doi.org/10.3262/ZP2106840

Kohrt, P., Haag, N. & Stanat, P. (2017). Kompetenzstufenbesetzungen im Fach Mathematik. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 140–152). Münster: Waxmann.

Kultusministerkonferenz. (2015). *Förderstrategie für leistungsstarke Schülerinnen und Schüler. Beschluss der Kultusministerkonferenz vom 11.06.2015*.

Lintorf, K. (2012). Messtheoretische Güte von Schulnoten. In K. Lintorf (Hrsg.), *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale* (S. 37–66). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94339-8_3

Lüders, M. (2001). Probleme von Lehrerinnen und Lehrern mit der Beurteilung von Schülerleistungen. *Zeitschrift für Erziehungswissenschaft*, *4*(3), 457–474. https://doi.org/10.1007/s11618-001-0047-6

Maaz, K., Baeriswyl, F. & Trautwein, U. (2013). „Herkunft zensiert?“ Leistungsdiagnostik und soziale Ungleichheiten in der Schule ["Origin graded?" Performance Diagnostics and Social Inequalities at School]. In D. Deißner (Hrsg.), *Chancen bilden. Wege zu einer gerechteren Bildung - ein internationaler Erfahrungsaustausch* (S. 185–188). Wiesbaden: Springer Fachmedien Wiesbaden.

Neuendorf, C., Kuhl, P. & Jansen, M. (2017). Leistungsstarke Schülerinnen und Schüler in Deutschland [High-achieving students in Germany]. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 317–334). Münster: Waxmann.

Pant, H. A., Böhme, K., Stanat, P., Schipolowski, S. & Köller, O. (2019). Kompetenzstufenmodelle für das Fach Mathematik und für die naturwissenschaftlichen Fächer. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich & S. Henschel (Hrsg.), *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (S. 51–98). Münster, New York: Waxmann.

Patry, J.-L. (2013). Beyond multiple methods: Critical multiplism on all levels. *International Journal of Multiple Research Approaches*, *7*(1), 50–65. https://doi.org/10.5172/mra.2013.7.1.50

Pelkner, A.-K. & Boehnke, K. (2003). Streber als Leistungsverweigerer? Projektidee und erstes Datenmaterial einer Studie zu mathematischen Schulleistungen [Nerds as refusers of performance?]. *Zeitschrift für Erziehungswissenschaft*, *6*(1), 106–125.

Pelkner, A.-K., Günther, R. & Boehnke, K. (2002). Die Angst vor sozialer Ausgrenzung als leistungshemmender Faktor? Zum Stellenwert guter mathematischer Schulleistungen unter Gleichaltrigen. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen.* Weinheim: Beltz. https://doi.org/10.25656/01:23587

PISA 2009 Technical Report. (2012). OECD. https://doi.org/10.1787/9789264167872-en

Reinhold, F., Reiss, K., Diedrich, J., Hofer, S. & Heinze, A. (2019). Mathematische Kompetenz in PISA 2018 - aktueller Stand und Entwicklung. In K. Reiss, M. Weis & E. Klieme (Hrsg.), *PISA 2018. Grundbildung im internationalen Vergleich* (S. 187–210).

Rheinberg, F. (2008). Bezugsnormen und die Beurteilung von Lernleistung. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 178–186). Göttingen: Hogrefe.

Rohrer, J. M. (2021, 7. März). *Mülltiverse analysis. The 100% CI*. Verfügbar unter: http:// www.the100.ci/2021/03/07/mulltiverse-analysis/

Schiepe-Tiska, A., Rönnebeck, S. & Neumann, K. (2019). Naturwissenschaftliche Kompetenz in PISA 2018 - aktueller Stand, Veränderungen und Implikationen für die naturwissenschaftliche Bildung in Deutschland. In K. Reiss, M. Weis & E. Klieme (Hrsg.), *PISA 2018. Grundbildung im internationalen Vergleich* (S. 211–240).

Schipolowski, S., Stanat, P., Böhme, K., Haag, N., Sachse, K. A., Hoffmann, L. et al. (2016). Der Blick in die Länder. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrsg.), *IQB-*

Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich (S. 171–330). Münster: Waxmann.

Shadish, W. R. (1986). Planned critical multiplism: Some elaborations. *Behavioral Assessment*, *8*(1), 75–103.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, *1993*(60), 13–57. https://doi.org/10.1002/ev.1660

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S. & Heschel, S. (Hrsg.). (2019). *IQB Bildungstrend 2018. Mathematisch und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich. Zusatzmaterialien*. Münster, New York: Waxmann.

Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Weigand, G. (2020). "Leistung macht Schule" - Eine Einführung. In G. Weigand, C. Fischer, F. Käpnick, C. Perleth, F. Preckel, M. Vock et al. (Hrsg.), *Leistung macht Schule. Förderung leistungsstarker und potenziell besonders leistungsfähiger Schülerinnen und Schüler* (Pädagogik, 1. Auflage, S. 13–22). Weinheim: Beltz.

Weirich, S., Wittig, J. & Stanat, P. (2017). Kompetenzstufenbesetzungen im Fach Deutsch. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 129–139). Münster: Waxmann.

Weis, M., Doroganova, A., Hahnel, C., Becker-Mrotzek, M., Lindauer, T., Artelt, C. et al. (2019). Lesekompetenz in PISA 2018 - Ergebnisse in einer digitalen Welt. In K. Reiss, M. Weis & E. Klieme (Hrsg.), *PISA 2018. Grundbildung im internationalen Vergleich* (S. 47–80).

Weis, M. & Reiss, K. (2019). PISA 2018 - Ziele und Inhalte der Studie. In K. Reiss, M. Weis & E. Klieme (Hrsg.), *PISA 2018. Grundbildung im internationalen Vergleich* (S. 13–20).

Wentzel, K. R., Jablansky, S. & Scalise, N. R. (2021). Peer social acceptance and academic achievement: A meta-analytic study. *Journal of Educational Psychology*, *113*(1), 157–180. https://doi.org/10.1037/edu0000468

Young, C. & Holsteen, K. (2017). Model Uncertainty and Robustness. *Sociological Methods & Research*, *46*(1), 3–40. https://doi.org/10.1177/0049124115610347

Ziegler, A. (2018). Hochbegabte, Begabtenförderung und Bildung. In R. Tippelt & B. Schmidt-Hertha (Hrsg.), *Handbuch Bildungsforschung* (S. 1279–1296). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-19981-8_57

3

Wie entwickeln sich Leistungsstarke?

1. Die differenzielle Leistungsentwicklung leistungsstarker Schülerinnen und Schüler

1.1. Theoretischer Hintergrund

Der zweite Beitrag dieser Dissertation handelt von der Leistungsentwicklung leistungsstarker Schülerinnen und Schüler. Diese wird untersucht, indem unterschiedliche Leistungsgruppen kontrastiert werden. Die differenzielle Entwicklung verschieden leistungsstarker Gruppen wird meist unter dem Paradigma des Matthäus-Effekts untersucht. Dabei zeigt ein Überblick über die bisherige Literatur, dass die Verwendung des Begriffs einem Wandel unterlag. Der Matthäuseffekt bezieht sich begrifflich auf die Bibel, wo es an unterschiedlichen Stellen heißt: „Denn wer da hat, dem wird gegeben, daß er die Fülle habe; wer aber nicht hat, von dem wird auch das genommen was er hat“ (*Lutherbibel,* 1912, Mt 13:12; 25:29). In der Forschung wurde der Ausdruck erstmalig prominent von Richard Merton und Harriet Zuckerman genutzt, die damit einen Mechanismus beschrieben, bei dem in der Wissenschaft Forschungsarbeiten stärker denen zugeschrieben werden, die bereits berühmter sind. Unterschiedliche Ausgangsleistungen bedingen so einen kumulativen Vorteil für diese Forschenden. Dieses ursprüngliche Verständnis war also das von Fehlattribution von Anerkennung in der Wissenschaft (Merton, 1968; Zuckerman, 2010). Das Konzept wurde in der Folge in unterschiedlichen Forschungsbereichen übernommen und auf eine Vielzahl von Themen adaptiert. Dabei erfuhr es vielfach eine Bedeutungsänderung. Zuckerman (2010) zeichnet diese Entwicklung des Begriffs nach. Innerhalb der Bildungsforschung wurde der Begriff zunächst von Walberg und Tsai (1983) verwendet. Sie untersuchten Prädiktoren für Testleistungen von jungen Erwachsenen und stellten fest, dass die vergangenen Bildungserfahrungen einen größeren Einfluss hatten als motivationale Variablen und aktuelle Lernaktivitäten. Sie stellten die Hypothese auf, dass vergangenes Lernen dazu führt, dass künftige Lernprozesse effizienter ablaufen können. Darauf aufbauend entwickelte Stanovich (1986) die Matthäus-Effekt-Hypothese der Leseentwicklung, welche die Grundlage für viele weitere Forschungsarbeiten bildete. Stanovich nahm an, dass zwischen Lesekompetenzentwicklung und Lesepraxis ein reziproker Zusammenhang besteht, welcher über Motivation vermittelt wird und zu einer Aufwärtsspirale führt. Das heißt, Kinder, welche bereits einen Vorsprung in der Lesekompetenz zu Beginn der Schule aufweisen, seien motivierter zu lesen und lesen daher mehr, was die Kompetenzentwicklung günstig beeinflusse, sodass die weitere Entwicklung im Vergleich mit anfänglich weniger kompetenten Mitschülerinnen und Mitschülern beschleunigt sei und ein sich progressiv vergrößernder Abstand zwischen diesen Schülergruppen bestehe. Es entstünde dadurch ein kumulativer Vorteil für diese Kinder.

DiPrete und Eirich (2006) bereiteten den Boden für weitere Untersuchungen des Matthäus-Effekts, indem sie das Konzept des kumulativen Vorteils formalisierten. Sie buchstabierten

unterschiedliche mögliche statistische Mechanismen und Verständnisse des Konzepts des kumulativen Vorteils aus und ordneten den Begriff des Matthäus-Effekts ein. Sie unterschieden zunächst zwischen pfadabhängigen und statusabhängigen Prozessen kumulativen Vorteils. Bei pfadabhängigen Prozessen, welche Sie als kumulativer Vorteil im engeren Sinne bezeichneten, hängt der weitere Zuwachs an Wissen und Kompetenz ausschließlich vom vorherigem Kompetenzniveau ab, beziehungsweise "future accumulation depends on current accumulation" (DiPrete & Eirich, 2006, S. 375). Diese pfadabhängigen Prozesse können sowohl linear (bei konstanten Wachstumsraten) als auch nichtlinear, also exponentiell beschleunigt (bei Wachstumsraten, die in Abhängigkeit von der aktuellen Leistung variieren) verlaufen.

Bei statusabhängigen Prozessen kann sich eine Leistungsschere öffnen, wenn verschiedene Statusgruppen unterschiedliche Wachstumsraten aufweisen, also andere stabile endogene und exogene Faktoren einen Einfluss auf die Entwicklungsgeschwindigkeit haben. Dies können beispielsweise Behinderungen (Morgan, Farkas & Hibel, 2008), familiäre Risikofaktoren (niedriger SES: Crawford, Macmillan & Vignoles, 2016) oder unterschiedliche Lern- und Entwicklungsmöglichkeiten (Schulformen: Pfost, Karing, Lorenz & Artelt, 2010) sein. Diese Prozesse sehen insbesondere dann wie Matthäuseffekte aus, wenn die Statusgruppen bereits zu Beginn unterschiedliche Ausgangswerte aufweisen.

Bisherige Forschung hat Prozesse kumulativen Vorteils in der Kompetenzentwicklung folglich entweder unter dem Paradigma der sich vergrößernden Disparitäten unterschiedlicher Gruppen oder als "echten" Matthäuseffekt, also eines (bei gleichem Lernangebot) differenziellen Wachstums von zu Beginn unterschiedlich leistungsstarken Gruppen (Ceci & Papierno, 2005; Walberg & Tsai, 1983). Zu ersterem wird hier als Beispiel der Schulformeffekt hervorgehoben. Forschung hat gezeigt, dass unterschiedliche Leistungsentwicklungen zwischen Schülerinnen und Schülern oft (aber nicht durchgängig, siehe Baumert, Becker, Neumann & Nikolova, 2009) durch die besuchte Schulform erklärt werden können (Kocaj, Kuhl, Kroth, Pant & Stanat, 2014; Pfost et al., 2010). Da die besuchte Schulform jedoch gleichzeitig mit der Ausgangsleistung korreliert ist, würde sich in der Gesamtbetrachtung ein kumulativer Vorteil für leistungsstärkere Schülerinnen und Schüler ergeben (Pfost et al., 2010).

Die Leistungsentwicklung von Schülerinnen und Schülern unterschiedlicher Ausgangslage wurde bislang meist an Grundschulen und meist für die Leseentwicklung untersucht. Beispielweise fassten Pfost, Hattie, Dörfler und Artelt (2014) vergangene Forschung zusammen und stellten fest, dass bisher kein klares Bild existiert und keines der möglichen Muster – wachsende Leistungsunterschiede, kompensatorische Prozesse oder parallele Entwicklung – dominant war.

Sie schließen daraus, dass Moderatoren, welche die unterschiedlichen Ergebnisse erklären können, untersucht werden sollten. In ihrer Meta-analyse wurden die Einflüsse verschiedener Studieneigenschaften auf die Ergebnisse überprüft: keinen Effekt konnten sie für die Moderatoren Sprache, Altersgruppe und Untersuchungsmethode nachweisen. Hingegen gab es Hinweise auf einen Einfluss des untersuchten Konstrukts (deskriptive Ergebnisse unterstützten eher kompensatorische Prozesse bei Lesekompetenz/-verständnis und eher stabile bzw. Matthäuseffekte für Lesegeschwindigkeit) und auf einen Einfluss der Qualität des eingesetzten Testes (Tests ohne Decken-/Bodeneffekte zeigten ausgeglichene Ergebnisse; Tests mit Decken-/Bodeneffekten führten eher zu kompensatorischen Prozessen; Tests mit geringerer Reliabilität waren ebenfalls eher bei Studien mit kompensatorischem Ergebnismuster zu finden. Studien ohne Angabe der Reliabilität zeigten eher ein Matthäus-Effekt-Muster).

1.2. Ableitung der Forschungsfrage und methodisches Vorgehen

In Artikel II wird daher die Frage untersucht, ob leistungsstarke Schülerinnen und Schüler in Lesen oder Mathematik einen kumulativen Vorteil gegenüber ihren weniger leistungsstarken Peers besitzen, das heißt, ob anfängliche Unterschiede sich vergrößern. Um die Schulform konstant zu halten, wurden ausschließlich Schülerinnen und Schüler an Gymnasien untersucht. Dies geschah, um den differenziellen Effekt gleicher Lernbedingungen und damit den Matthäuseffekt im engeren Sinne zu untersuchen. Das Gymnasium ist die Schulform, welche insbesondere dafür gedacht ist, leistungsstarke Schülerinnen und Schüler angemessen zu fördern. Daher wäre es durchaus plausibel, anzunehmen, dass die Leistungsspitze sich gegenüber Mitschülerinnen und Mitschülern im Laufe der Sekundarstufe weiter absetzt. Hingegen würde eine stärkere Homogenisierung der Leistungen dafür sprechen, dass die Schule insgesamt eher zur Förderung leistungsschwächerer Schülerinnen und Schüler beiträgt. Bisherige Untersuchungen zum Summer learning gap legen genau dies nahe: Leistungsschwächere lernten im Verlaufe des Schuljahres stärker hinzu als Leistungsstärkere, büßten diesen Lernfortschritt aber im Verlauf der Sommerferien wieder ein, wenn Leistungsstarke sich in ihrer Kompetenz konstant weiterentwickelten, während ihre Mitschülerinnen und Mitschüler wieder gegenüber den anderen zurückfielen (Alexander, Entwisle & Olson, 2001, 2007). Ob Unterschiede in den Wachstumsraten an Gymnasien gefunden werden können, einer Schulform, welche sich der Förderung der Leistungsspitze verschrieben hat, wird in der aktuellen Studie überprüft.

Im Artikel wird auf Grundlage von Daten der BiKS-Studie (Artelt, Blossfeld, Faust, Roßbach & Weinert, 2013) die Leistungsentwicklung von Schülerinnen und Schülern von Klassenstufe 5, also direkt nach dem Übergang auf die weiterführende Schule, bis Klassenstufe 9 nachgezeichnet. Dabei wird sowohl die Leistungsentwicklung in Deutsch als auch in Mathematik betrachtet, da

davon ausgegangen werden kann, dass es Unterschiede zwischen den Fächern gibt. Während das Lesen durch Lesegewohnheiten in der Freizeit auch außerschulisch gefördert wird, so ist das Vermittlungsmonopol bei Mathematik stärker in der Schule zu verorten (Prenzel, Reiss & Hasselhorn, 2010).

Zur Untersuchung wird, wie von Bast und Reitsma (1998) vorgeschlagen, die Methode der Wachstumskurvenanalyse eingesetzt. Diese Art von Strukturgleichungsmodellen bietet die Möglichkeit, Entwicklungsverläufe über mehrere Messzeitpunkte zu modellieren. Eine Möglichkeit, die Abhängigkeit der Leistungsentwicklung vom Ausgangswert zu quantifizieren, stellt die Intercept-Slope-Korrelation dar. Ist diese positiv, bedeutet dies, dass leistungsstärkere Schülerinnen und Schüler eine höhere Wachstumsrate haben. Ist sie negativ, spricht dies eher für kompensatorische Prozesse. Allerdings könnte es sein, dass die Unterschiede in den Wachstumsraten nicht proportional zum Ausgangswert sind. Um dennoch Aussagen auch über Extremgruppen treffen zu können, wird eine Multigruppenanalyse durchgeführt. So können nicht-lineare Effekte der Ausgangsleistung auf die weitere Leistungsentwicklung untersucht werden. Weiterhin wird die Frage gestellt, ob die Leistungsentwicklung über den Verlauf der Sekundarstufe I hinweg linear verläuft, oder ob die Entwicklung innerhalb der Gruppen sich mit der Zeit verändert. Hierfür werden unterschiedliche Steigungsparameter für unterschiedliche Zeitabschnitte geschätzt. Somit entsteht ein genaueres Bild der Leistungsentwicklung von Gruppen verschiedener Ausgangslage. Um die Robustheit der Ergebnisse über verschiedene Operationalisierungen von Leistungsstärke einschätzen zu können, wird schließlich eine Multiversumsanalyse durchgeführt, welche den Einfluss unterschiedlicher Cut-off-Werte für die Gruppenbildung auf den Unterschied in Wachstumsraten zwischen den Gruppen sowie die Sensitivität des Effekts gegenüber dem Testinstrument visualisiert.

II

Competence development of high achievers within the highest track in German secondary school: Evidence for Matthew effects or compensation?

Claudia Neuendorf

Malte Jansen

Poldi Kuhl

Neuendorf, C., Jansen, M. & Kuhl, P. (2020). Competence development of high achievers within the highest track in German secondary school: Evidence for Matthew effects or compensation? Learning and Individual Differences, 77, 101816. https://doi.org/10.1016/j.lindif.2019.101816.

Abstract

The Matthew effect hypothesis of academic development predicts that students with higher initial achievement will develop further skills at a faster rate resulting in cumulative advantages. Prior research has focused on the development of reading competence in primary school. To extend this research, we used a sample of N = 1,010 German students in Grades 5 to 9 to compare the development of reading and mathematics skills between high achieving high-track secondary school students and their peers to clarify whether rates of academic development differ between these groups. Using latent growth curve modeling, we found a pattern of compensation in both domains—that is, the achievement gap became smaller and this was the case particularly in the early grades of secondary school. Thus, our results provide no evidence for the existence of Matthew effects in reading and mathematics in lower secondary school.

Introduction

Children enter school with different skills and levels of prior knowledge. For teachers, this heterogeneity represents a challenge, because they are expected to support all students realize their full potential while considering that these differences in prior knowledge and skills have an impact on their further learning and academic development. Teachers facing the dilemma of how to distribute their resources might endorse the opinion that high achievers need no specialized promotion because they are already advantaged and will continue to succeed anyway (Bannister 2016; Brighton, Hertberg, Moon, Tomlinson, & Callahan, 2005). However, to date, it is unclear whether such students continually outperform other students or whether their advantages level off across the school career. Thus, research should investigate, in how far schools currently succeed in promoting the advancement of high achievers compared to other students.

Until now, the academic development of students at different levels of initial achievement has often been studied from the viewpoint of cumulative advantage (e.g. Baumert, Nagy, & Lehmann, 2012; Cunningham & Stanovich, 1997; Shaywitz, Holford, Holahan, Fletcher, Stuebing, Francis, & Shaywitz, 1995; Walberg & Tsai, 1983). This perspective stresses that success facilitates further success, and small advantages in an early stage of a growth process grow progressively larger (DiPrete & Eirich, 2006). For the academic domain, it suggests that children with a head start in a certain subject or domain will continually make more progress than their peers, given the same instruction. This process is also referred to as Matthew effect (Ceci & Papierno, 2004; for a discussion on the usage of the term, see Zuckerman, 2011).

To our knowledge, empirical research on the Matthew effect has mostly focused on reading in primary school contexts. Research focusing on secondary school or research looking at different subjects is more scarce. Specifically, research on achievement gaps in secondary school mostly looked at between track differences in learning gains (e.g. Becker et al., 2014; Dockx, de Fraine, & Vandecandelaere, 2018; Kulik & Kulik, 1982; Pfost & Artelt, 2013). In contrast, this paper will describe differences in academic growth *between* high achievers and their peers *within* the highest track of secondary school.

Therefore, with the current study, we aimed to extend previous research by clarifying whether either the cumulative advantage hypothesis (i.e., the Matthew effect) or the compensation hypothesis, which predicts that achievement gaps would decrease over time, can be substantiated for upper track secondary school students' academic development in reading and mathematics or whether neither of them apply and parallel development for different achievement groups can be observed. In particular, we were interested in determining whether reading and mathematics would

show similar result patterns. By using a longitudinal German data set with yearly measures from Grades 5 to 9 in reading and from Grades 5 to 8 in mathematics, we covered an important phase in students' academic and personal development starting right after their transition to secondary school.

Academic achievement gap: Cumulative advantage versus compensation

When studying differences in growth between students, one of two competing hypotheses can be put forward: the cumulative advantage (or Matthew effect) hypothesis or the compensation hypothesis. Before examining empirical evidence, in the following, the assumptions behind either of the two are being discussed.

The concept of cumulative advantage with its different interpretations has been specified by DiPrete and Eirich (2006) and Baumert et al. (2012). They have distinguished between *status dependent processes of cumulative advantage* and *path dependent processes of cumulative advantage*. While the former are the result of the differential impact of individual characteristics (e.g., socioeconomic status, race, gender, cognitive abilities) on learning gains and lead to increasing differences between status groups, the latter occur solely based on previous achievement level which enables high achievers to learn more from the same instruction. These path dependent processes of cumulative advantage are the focus of the current paper. It should be possible to find empirical evidence for them in patterns of diverging cohort trajectories between initial high achievers and others.

A domain specific model predicting Matthew effects in reading has been developed by Stanovich (1986). He assumed that the accelerated pace of reading development in high achievers would be mediated by higher reading motivation and reading practice in this group. This mechanism is assumed to work in particular in the beginning of reading skill acquisition and has been studied extensively in primary school (Pfost, Hattie, Dörfler, & Artelt., 2014).

Whereas such an explicit model of differential academic development for different achievement groups does not exist for mathematics, the idea of Matthew effects in mathematics is very plausible if the cumulative nature of mathematical skills is acknowledged such that from basic to more advanced skills, each step of learning is based on prior knowledge and skills. Consequently, the assumption of cumulative advantage has been suggested not only for reading but also for mathematics, thus implying that children who initially show higher achievement in mathematics also show higher rates of growth than children with less prior knowledge in mathematics (Claessens & Engel, 2013; Duncan et al., 2007).

Whereas research on the Matthew effect has so far been mainly concerned with primary school development, the self-reinforcing mechanism resulting in cumulative advantage could in principle also be valid for secondary school. That is, higher achievement might lead to the more (effective) use of learning opportunities and to faster rates of learning throughout students' school careers.

However, for most skills taught in school, the acceleration of student learning may not subsist. It has been argued that some skills have a natural ceiling that is reached earlier by some students, and then no or only limited further skill advancement can occur (Baumert et al., 2012). Thus, Matthew effects might not reflect genuinely different rates of growth but merely different timing in the onset of a developmental process and might be followed by compensation processes when "late bloomers" catch up (developmental-lag model; Pfost et al., 2014). As a result, secondary school might also be a time when students with a lower ability level close the achievement gap between themselves and others. This could be seen in compensatory development.

However, additional factors that are not present in primary school affect academic development in secondary school and may increase or decrease the Matthew effect. In Germany, where the current study was located, students enter into an institutionally tracked education system after primary school. The different tracks provide different learning environments with different curricula, different teaching cultures, and different teacher training study programs, but also with different compositions of students (Baumert, Stanat, & Watermann, 2006; Pfost & Artelt, 2013). Academic development in secondary school has mostly been studied either with a focus on the effects of tracking in general or on the effects of ability grouping (Baumert, Becker, Neumann, & Nikolova, 2009; Kulik & Kulik, 1982; Maaz, Trautwein, Lüdtke, & Baumert, 2008; Pfost & Artelt, 2013; Pfost, Karing, Lorenz, & Artelt, 2010). However, so far no research has focused on the difference between high achievers and other students within the highest track of secondary school.

Various mechanisms within the highest track of secondary school could lead to a continuing pattern of cumulative advantage in initial high achievers. Among these are effects of instructional design in high track schools and compositional effects, which we elaborate on in the following.

Instruction and differences in achievement growth

The teaching strategies employed in high-track schools are characterized by cognitively activating and demanding instruction methods (Kunter et al., 2005; Slavin, 1990). However, the effectiveness of different teaching strategies has been found to vary with students' ability level (Caro, Lenkeit, & Kyriakides, 2016; Snow & Lohman, 1984). Instructional design in the high-track schools might thereby be more tailored to higher aptitude students, for example, such that lessons are less tightly structured or less feedback is given (Baumert, Maaz, Stanat, & Watermann, 2009; Fyfe, Rittle-

Johnson, & DeCaro, 2012; Jennek, Gronostaj, & Vock, 2019). This learning environment might be especially well suited to high achievers' cognitive and motivational preconditions, possibly leading to an acceleration of their future development compared with their peers, thus supporting the Matthew effect hypothesis.

But, although frequently promoted, differentiation practices, which should enable all students to progress at high rates, seem to be rather rare in classroom practice (Archambault et al., 1993, Tomlinson et al., 2003). Where differentiation is employed, it is often implemented with a focus on students who are struggling to learn, rather than on more advanced students (Brighton, Hertberg, Moon, Tomlinson & Callahan, 2005; Plucker & Callahan, 2014). This might be due to limited resources or skills of teachers to implement differentiated teaching and to provide appropriately challenging tasks for high achievers (Brighton et al., 2005). But it might also be due to teacher beliefs and attitudes (McCoach & Siegle, 2007; Tomlinson et al., 2003). Thus, higher gains for lower achieving students or a pattern of compensation may result from teachers aiming for homogeneity in their classrooms and devoting more time to lower achieving students.

Classroom composition and differences in achievement growth

After primary school, the most advanced students have the opportunity to transfer to the highest track secondary schools. Thus, the achievement standard within the new classes is raised compared to primary school. The opportunity to compare their abilities and achievement to students who match them more closely has been found to be a motivational factor for those performing at the highest levels (Li & Adamson, 1992; Phillips & Lindsay, 2006; Udvari & Schneider, 2000). It might stimulate them to accelerate the pace of their academic development and result in cumulative advantage.

But, again, there are aspects related to class composition which favor lower achievers within a higher achieving class. In particular, the composition of the student body in high-track schools has been found to be an effective resource for academic development (Dar & Resh, 1986), even though this often comes at the price of a lowered academic self-concept as a result of upward social comparisons (Trautwein, Lüdtke, Marsh, & Nagy, 2009). A number of studies have shown that the positive impact of being in a high-achieving class on children's academic development is higher for children with lower achievement levels (Dar & Resh, 1986; Kiss, 2013; Linchevski & Kutscher, 1998; Ma, 2005). However, whether this effect offsets the positive effect of higher initial competencies on subsequent performance and can therefore cause compensation effects is questionable (Kiss, 2013).

Thus, two central aspects students are confronted with after transitioning to secondary school—compositional effects as well as teaching methods within the highest track—might favor high achievers or low achievers and no clear prediction can be made as to whether a pattern of growing differences or a compensatory pattern can be expected.

Previous research findings

Prior research on path-dependent processes of cumulative advantage in reading has produced mixed results. In their meta-analysis of Matthew effects in reading development in primary school, Pfost et al. (2014) concluded: "We did not find strong support for the general validity of a pattern of widening achievement differences or for a pattern of decreasing achievement differences in reading" (Pfost et al., 2014, p. 203). However, this meta-analysis, like many studies in the field focused on primary school contexts. Some studies also included the later years of primary school, that is, Grades 4 to 6. Examining these studies' results for the later years, one can either use intercept-slope correlations of growth curve models or calculate effects sizes of differences in reading growth between high and low achieving groups. Intercept-slope correlations in reading were between $r = -.07$ and $r = -.63$ (Baumert et al., 2012; Morgan, Farkas & Wu, 2011; Shin, Davison, Long, Chan & Heistad, 2013). Effect sizes of between-group differences in growth amounted to between $d = -0.01$ and $d = -0.21$ (Aarnoutse, van Leeuwe, Voeten & Oud, 2001; Rambo-Hernandez & McCoach, 2015). All of these results were in favor of lower achieving students. On the contrary, Scammacca, Fall, Capin, Roberts & Swanson (2019) reported higher growth rates of high achievers compared to low achievers, with an effect-size difference of $d = 0.11$ across Grade 5. However, these studies differed in which groups they compared: While Rambo-Hernandez and McCoach (2015) compared the top 2 % with an average achieving group (16^{th}-84^{th} percentile), Aarnoutse and von Leeuwe (2000) and Aarnoutse et al. (2001) split the sample to form a group of 30 % lower achieving and 70 % higher achieving students and Scammacca et al. (2019) compared the lowest and the highest quartile.

For the development of mathematical skills, less research has been conducted, but studies comparing growth rates between groups of children with different early mathematics skills have also revealed mixed results. In their review, Nelson and Powell (2018) point out, that most studies on the longitudinal development of different ability students in math are conducted in primary school. Those studies that go beyond third grade have mostly found evidence for a Matthew effect (Morgan, Farkas & Wu, 2009; Morgan, et al., 2011; Lu, 2016), with intercept-slope correlations between $r = .36$ and $r = .46$. Shin et al. (2013) showed intercept-slope-correlations as low as $r = .10$ across Grades 4-7 and Ma (2005) reported very high intercept-slope correlations between $r = .55$ and $r = .99$. An exception is Scammacca (2019): In this study, the difference in growth

between high achievers (top 25%) and low achievers (bottom 25%) across Grade 5 was d = -0.26 in favor of lower achievers. Thus, while for reading, most studies have found compensation effects, studies on the development of mathematic skills predominantly found Matthew effects.

What is noteworthy about the previous research in both domains is the dominant focus on trajectories of poorly performing or low ability students (Aarnoutse & van Leeuwe, 2000; Claessens & Engel, 2013; Jordan, Hanich, & Kaplan, 2003; Kühn, Sachse, & Suchodoletz, 2015; Pfost et al., 2012; Shin et al., 2013; Skibbe et al., 2008) and the lack of studies focusing on high achievers (for an exception, see Rambo-Hernandez & McCoach, 2015).

The present study

Our study focuses on the question of who learns more in the highest track of secondary school—initial high achievers or their relatively lower achieving peers. Reviewing the current literature on the topic, we identified certain gaps we are aiming to address. Fistly, the vast majority of studies have investigated academic development of different achievement groups in primary school contexts and secondary school has less often been the period under investigation. Secondly, when previous studies contrasted different achievement groups, they mostly focused on low-achieving children (Francis et al., 1996; Jordan et al., 2003; Pfost et al., 2012; Shin et al., 2013; Skibbe et al., 2008). High achievers have been explicitly considered in few studies (for example, Rambo-Hernandez & McCoach, 2015). Thirdly, academic development in secondary school in general and the development of the achievement gap between high achievers and others might be nonlinear. But, nonlinear trajectories have not been considered in many studies (see Baumert et al., 2012; Pfost et al., 2012; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996, Shin et al., 2013, for exceptions). Furthermore, not many any studies have analyzed both reading and mathematics skills simultaneously, making results comparable between different domains (see Baumert et al., 2012, for an exception).

Thus, our study aimed to examine learning trajectories in reading and mathematics in secondary school and to evaluate the evidence for differential growth patterns of initially high-achieving students compared with other students. Based on the theoretical assumptions and the previous research described above, three conflicting hypotheses can be formulated regarding differences between high achievers and their peers in reading and mathematics: First, the trajectory of high achievers compared to their peers may be characterized by a pattern of cumulative advantage (i.e., steeper growth for initially high-achieving students compared with students with lower achievement) as a result of their higher prior skills and knowledge as well as the more stimulating learning environment in secondary school (H1.a). On the contrary, other students might also catch

up— because the compositional effect of being in a high-achieving class favors them, because teachers might invest more resources in them, or simply because they follow higher achievers on a non-linear trajectory —leading to a pattern of compensation (H1.b). Finally, these processes might cancel each other out, leading to parallel achievement trajectories (H1.c). Given the conflicting previous evidence, we did not a priori favor one of these hypotheses.

Second, we aimed to determine whether growth rates (and group differences in growth rates) change over time, indicating nonlinear developments. Based on the previous research (Baumert et al., 2012; Pfost et al., 2012; Francis et al., 1996, Shin et al., 2013), we expected that the academic development of students in reading and mathematics follows a non-linear shape across secondary school. Specifically, we expected the growth rates of high achievers and their peers to show the largest differences in the early years of secondary school (H2). During this time, within the newly formed learning groups social and cognitive adjustment processes take place. Moreover, teachers might strive for a more homogeneous level of knowledge and skills in their new classes.

Third, we were interested in whether development in reading and mathematics would show similar patterns. While there are some processes which would affect development in both subjects simultaneously (newly formed classes in fifth grade, deceleration of academic development in general), others might be subject specific (nature of the skills to be learned, learning opportunities inside and outside of school). Because of the cumulative structure of the mathematics curriculum, the increasing complexity of the learning content and the confinement of mathematics learning to the mathematics classroom, while reading practice is almost ubiquitous (in different school subjects as well as outside school), we expected a larger difference in trajectories in favor of the high achievers in mathematics compared to reading (H3).

We used a German panel study that collected achievement data in reading and mathematics at five and four measurement points, respectively, with the first measurement point taking place at the beginning of secondary school. We addressed our research questions in the structural equation modeling framework by modeling paths of academic development using latent growth curve models and applied different modeling approaches to check our results for robustness. Further, we paid special attention to the issues of measurement invariance across measurement points by linking the scales longitudinally.

Method

Study design

We used data from the BiKS study (Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vorschul- und Schulalter [Educational processes, competence

development and selection decisions in preschool and school age] (Artelt, Blossfeld, Faust, Roßbach, & Weinert, 2013), which was conducted between 2006 and 2014 in the two German federal states of Bavaria and Hesse. In this study, approximately 2,400 children were followed from third grade to ninth grade. In Bavaria and Hesse, as in most German federal states, primary school comprises the first four years of schooling, after which children transfer to the tracked secondary school system (for a brief overview of the German tracking system, see Supplement A). In the study, students were tested annually with a selection of ability and achievement tests, and students, teachers, and parents also answered questionnaires. In order to test school children, researchers had to obtain approval by the relevant Ministries of Education, which entails ethical approval as well as compliance with data protection guidelines. A detailed description of the study can be found in Lorenz, Schmitt, Lehrl, Mudiappa, and Rossbach (2013).

Analysis sample

To address the question of whether patterns of cumulative advantage for high-achieving students can be found in high-track secondary school either in reading or in mathematics, we used data from the beginning of secondary school in fifth grade to ninth grade and restricted our sample to students attending *Gymnasium* (i.e., the highest secondary school track in Germany). About 53%

Table II-1

Descriptive Statistics

	T1 (Grade 5)	T2 (Grade 6)	T3 (Grade 7)	T4 (Grade 8)	T5 (Grade 9)
N (reading)	1010	855	797	369	334
N (math)	1010	858	761	361	-
$M_{reading}(SD)$	0.00 (0.64)	0.71 (0.86)	1.00 (1.10)	1.68 (0.91)	1.9 (0.81)
$M_{math}(SD)$	-0.01 (0.81)	0.72 (1.00)	1.17 (1.02)	1.59 (1.10)	-
Stability reading					
Cor (t_1)	-	0.44	0.39	0.36	0.38
Cor (t_2)		-	0.54	0.44	0.33
Cor (t_3)			-	0.52	0.46
Cor (t_4)				-	0.57
Stability mathematics					
Cor (t_1)	-	0.56	0.52	0.45	
Cor (t_2)		-	0.60	0.51	
Cor (t_3)			-	0.53	

Note. Descriptive statistics were based on WLE scores. The data from the first wave of measurement were scaled with the Conquest option constraints = cases, that is, the mean of the latent dimension was set to zero. The following waves of measurement were anchored to the scale of the first measurement occasion.

of the students in the total sample attended this school type. We excluded students who had repeated a grade level because their learning trajectories necessarily differed. This left us with a total sample of 1,010 students in our analysis sample. At the first measurement occasion, 928 students participated in reading, and 892 participated in mathematics. As can be seen in Table II-1, the number of participating students decreased substantially over the course of the study. Our approach for dealing with missing data is described in Section 2.4 below. Students' demographic characteristics in Grade 5 are given in Table II-2.

High achievers were operationally defined as children scoring at least 1 SD above the total sample mean at the first measurement occasion. Thus, 12.5% of the students in Grade 5 were initially high achievers in reading, and 15.6% were initially high achievers in mathematics. For high achievers, there is currently no standard definition in the literature. In his Differentiated Model of Giftedness and Talent 2.0, Gagné (2009) called the top 10% of achievers "(mildly) talented." In international large-scale assessments, students are grouped in proficiency levels. In the previous PISA cycle, 11.7% of German students in reading and 12.9% in mathematics scored in the two highest proficiency levels (OECD, 2016). By choosing a cut-off value of 1 SD above the mean, we aimed

Table II-2

Participant Demographic Information

	Total sample	Reading			Mathematics		
	M (*SD*)	High achievers *M* (*SD*)	Other students *M* (*SD*)		High achievers *M* (*SD*)	Other students *M* (*SD*)	
1. *N*	1010	157	853		159	851	
2. Gender (% male)	45.45	46.50	45.25		48.43	44.89	
3. Age (years)	10.90 (0.40)	10.89 (0.40)	10.90 (0.40)		10.81 (0.38)	10.92 (0.40)	*
4. Parental university entry certificate (%)	68.90	76.16	67.51		78.67	67.05	*
5. Socioeconomic status (HISEI)	58.05 (15.61)	61.63 (15.02)	57.38 (15.64)	*	62.08 (14.68)	57.29 (15.67)	*
6. Immigration background (%)	18.97	15.13	19.70		12.00	20.28	*
7. Reading at t1	0.00 (0.64)	1.04 (0.41)	-0.19 (0.46)	*	0.44 (0.68)	-0.08 (0.60)	*
8. Mathematics at t1	-0.01 (0.81)	0.51 (0.82)	-0.11 (0.78)	*	1.27 (0.51)	-0.25 (0.61)	*

Note. HISEI = highest household international socio-economic index of occupational status (Ganzeboom, Graaf, & Treiman, 1992), students with an immigration background had either one or two parents who were born abroad.
*$p < .01$.

to describe a comparable group of high achieving students which is of practical significance. Our analysis decision was made prior to analyzing the data, but as a validity check, we re-analyzed the data using different cut-offs and found the same pattern of results. These additional analyses are detailed in supplement B.

Instruments

The *reading assessments* comprised three to five short texts with related multiple-choice items. The test was developed by the BiKS research group with additional items of increasing difficulty added in each wave (DFG-Forschergruppe "Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter" [BiKS], n.d.; Pfost, Dörfler, & Artelt, 2013). Between 25 and 43 items were administered during each wave of measurement and all reliabilities were acceptable (Cronbach's $\alpha > .74$).

The *mathematics assessments* were compiled from items from existing tests (DEMAT5+; Marx & Opitz-Karig, 2005, DEMAT6+; Götz, Lingel, & Schneider, 2013, PALMA; Pekrun et al., 2003, TIMSS; Baumert et al., 1998). The item content covered story problems, arithmetic, and geometry. The test forms used in the consecutive waves of measurement comprised between 29 and 44 items, and internal consistencies ranged from $\alpha = .75$ to $\alpha = .81$ in all waves. In contrast to reading, mathematics skills were not assessed in Grade 9. Thus, only four points of measurement were available. All subsequent waves of measurement in both reading and mathematics shared a set of common items—so-called anchor items—so that vertical scaling methods could be applied (Kolen & Brennan, 2014).

Data analysis strategy

IRT scaling and equating

To obtain ability estimates that were comparable between measurement occasions, we used the item response theory framework (IRT; Embretson & Reise, 2000) and equated the tests across all waves of measurement using the fixed item parameter calibration method (Kim, 2006; Kolen & Brennan, 2014; von Davier & von Davier, 2004). Thus, each wave was calibrated by fixing the anchor items to the estimated values of the previous wave (for a discussion on the comparability of achievement scores, see Protopapas, Parrila, & Simos, 2016). To ensure parameter invariance, we used an approach often recommended in the literature (Fischer, 1995) and excluded items that functioned differently between adjacent waves (i.e., items that showed item parameter drift) from the pool of anchor items. To this end, we first tested for differential item functioning (DIF) in the anchor items between two measurement occasions. We did this by scaling a test that contained only the anchor items and by treating students at the first measurement occasion as one group and

students at the following measurement as the second group. DIF between the two groups was estimated and evaluated according to the ETS classification scheme (Zieky, 1993). Only items that showed no significant DIF were included as anchor items. Sixteen and eight linking items had to be excluded in reading and mathematics, respectively, leaving a minimum of four and a maximum of 20 items to be used to link two subsequent waves. Based on these selected anchor items, the achievement tests were scaled separately using a one-parameter logistic IRT model with the software ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) in combination with the R package eatModel (R Core Team, 2018; Weirich & Hecht, 2018). Weighted Likelihood Estimates (WLEs; Warm, 1989) were used as estimates of student ability. In accordance with Pfost et al. (2012), missing values for the first measurement occasion were imputed using the EM Algorithm (Dempster, Laird, & Rubin, 1977) as implemented in the R-package dlsem (Magrini, 2016). This approach ensured that all children could be placed in either the group of high achievers or in the group of other students. For reading, data had to be imputed for 82 students, whereas for mathematics, data had to be imputed for 118 students. These WLE scores were used in all of the following analyses, which were carried out in the structural equation modeling framework.

Structural equation models

Latent growth curve models (LGCMs) are one common approach that can be applied to identify Matthew effects, especially when scales can be linked across waves (Bast & Reitsma, 1997). In linear LGCMs, trajectories of construct development across several measurement points are described by (a) the initial level (intercept) and (b) a linear increase or decrease (slope). Both the intercept and the slope are allowed to vary across students. If a process of cumulative advantage takes place, growth should be larger for higher achieving students than for other students (indicated by a higher slope estimate for this group). If, on the contrary, the data supports the compensation hypothesis, initial high achievers should show less growth in competencies than their lower achieving peers (indicated by a lower slope estimate for this group).

We first estimated one growth curve for all students across the measurement points (see Figure II-1A, Model 1). We then tested for nonlinear development in achievement by additionally estimating a quadratic growth curve (see Figure II-1A, Model 2). In a second step, we estimated separate growth curves for high achievers and others within a multigroup framework (Model 3). We compared these growth curves across the two groups of different attainment when they began secondary school. Estimating a two-group model also enabled us to investigate the average curves of the two groups visually and to test the model parameters (e.g., the mean of the slope factor) for equality. Third, we wanted to check whether the difference in growth between the two groups stemmed from a certain period of development, and therefore, we used a piecewise modeling

approach (Chou, Yang, Pentz, & Hser, 2004). Within this model, separate linear terms were estimated for the first and the second part of the growth curve. This approach constitutes an alternative nonlinear model, and it thus allowed us to explicitly compare different segments of the growth curve between the two groups (see Figure II-1B; Model 4). These steps were conducted for reading and mathematics separately.

The growth curve analyses were conducted in MPlus v 7.11 (Muthen & Muthen, 1998-2011), and the *Full Information Maximum Likelihood* (FIML) method was used to estimate all models using the whole sample to best capture all relations between the variables that were analyzed. This model-based approach is considered superior to traditional approaches (e.g., listwise or pairwise deletion) because it offers unbiased parameter estimation under the Missing-at-Random (MAR) assumption and retains statistical power because all observations are used (Enders, 2010; Wothke, 2000). However, because of the considerable share of missing data after Wave 3, which was mainly related to the failure of parents to renew their informed consent to the study (Homuth, Schmitt, Lorenz, & Mann, 2017), we re-analyzed the data using a model accounting for data missing not at random (NMAR). These analyses and results are detailed in Supplement B.

Results

Descriptive statistics

Table II-1 shows descriptive statistics for reading and mathematics achievement at each measurement point, as well as the stabilities (i.e., the autocorrelations) across all waves. As expected, mean reading and mathematics scores increased across the five or four points of measurement as the students progressed through secondary school. The total gains were 2.97 SDs

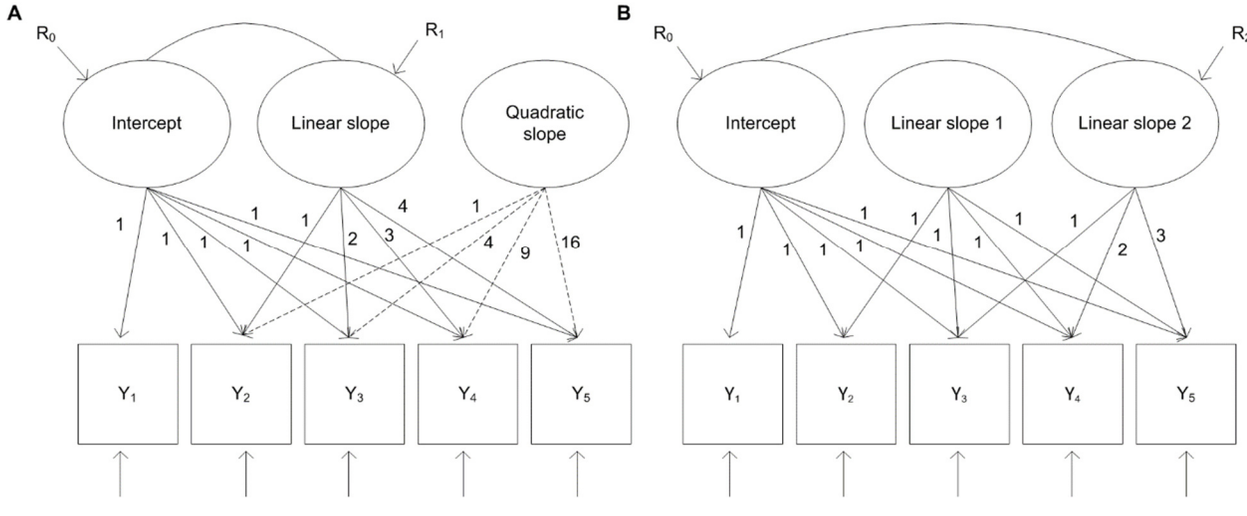


Figure II-1. Analysis model. (A) Model 1 (linear) and Models 2 and 3 (with additional quadratic growth factor); (B) Model 4 (piecewise growth model with two slopes).

based on the variance at Time 1 for reading and 1.95 SDs for mathematics. Stabilities were medium-sized, and the correlations between successive points of measurement were higher than they were between more distant points of measurement, thus showing the typical pattern for autocorrelations across time (Campbell & Kenny, 2003). The correlations between the first and last points of measurement were still $r_{r1\text{-}5} = .38$ and $r_{m1\text{-}4} = .45$ for reading and mathematics, respectively. These correlations indicate moderate rank-order stability in achievement but also substantial differences between students in skill development.

Finally, comparing raw growth rates of high achievers and other students across the time points, the differences were substantive, indicated by effect sizes ranging from $d = -1.08$ to $d = -1.40$ in reading and from $d = -.82$ to $d = -1.00$ in mathematics (see Table II-3). These negative effect sizes indicate that, for example, high achievers' standardized reading gain across the first period of measurement was 1.15 units smaller than the other group's gain, and their increase in mathematics achievement was 0.82 standardized units smaller than the other groups' increase.

General patterns of academic development

Reading achievement

As already suggested by the descriptive statistics, the LGC in Model 1 estimated an increase in reading ability (Table II-4). More specifically, the mean of the slope factor was 0.47—that is, on average, students' yearly achievement growth was equal to 0.47 units on the achievement metric at T1. The slope variance was also significantly greater than zero ($\sigma^2 = 0.02$, $p < .001$) indicating

Table II-3

Effect Sizes of Between-Group Differences in Growth

	T1-T2	T1-T3	T1-T4	T1-T5
Reading				
d	-1.15	-1.15	-1.40	-1.08
Mathematics				
d	-0.82	-1.00	-1.00	-

Note. Shown are the group differences (high achievers vs. other students) in gain scores between the first and the second (T1-T2), third (T1-T3), fourth (T1-T4) or fifth (T1-T5) measurement point. Effect sizes *d* are calculated using the two group's raw gain score and SD with a pooled SD adjusted for group size differences according to Hedges & Olkin (1985).

that students differed in their competence development. In order to determine whether the growth in reading followed a nonlinear pattern, a quadratic growth curve was estimated next (see Table II-4). A significant negative quadratic growth factor of -0.06 indicated that growth in reading development was slowing down across secondary school. Introducing a quadratic growth factor helped increase the model fit slightly (see Table II-5).

Mathematics achievement

Table II-6 shows results for the first two models of mathematics development. Similar to reading, the slope of 0.54 ($p < 0.01$) with its significant variance showed that there was overall growth but also significant differences between individuals in growth. In mathematics, there was also a significant negative quadratic growth, indicating a slowing down of learning processes. The quadratic model showed an improved model fit compared to the model without the quadratic term (see Table II-5).

Testing for differential competence development using multigroup models

To examine the different achievement trajectories for students with high versus lower achievement more closely than only based on intercept-slope correlations, we estimated multigroup growth curve models. Because the first models indicated that growth proceeded in a nonlinear fashion (see

Table II-4

Results of Latent Growth Curve Models (LGCMs) for Reading Development

	Model 1			Model 2		
	Est	(*SE*)		Est	(*SE*)	
Mean structure						
Intercept	0.05	(0.03)		0.01	(0.03)	
Linear slope	0.47	(0.01)	*	0.67	(0.04)	*
Quadratic slope				-0.06	(0.01)	*
Covariances						
Intercept – linear slope	0.01	(0.01)		0.01	(0.01)	
Variance structure						
Intercept	0.24	(0.02)	*	0.24	(0.02)	*
Linear slope	0.02	(0.00)	*	0.02	(0.00)	*
Quadratic slope				0.00	(0.00)	

Note. Model 1 = linear growth curve model; Model 2 = quadratic growth curve model.
*$p < .01$.

Table II-5

Fit Statistics for all Models

	CFI	RMSEA	SRMR	AIC	BIC
Reading					
Model 1	0.81	0.12	0.11	7741.95	7791.13
Model 2	0.87	0.11	0.13	7671.51	7725.61
Model 3	0.76	0.14	0.16	7021.11	7129.30
Model 4	0.85	0.11	0.12	6965.99	7084.02
Mathematics					
Model 1	0.93	0.11	0.06	7397.38	7441.64
Model 2	0.99	0.05	0.04	7331.02	7380.20
Model 3	0.96	0.08	0.07	6714.65	6813.01
Model 4	0.82	0.14	0.13	6803.63	6882.31

Note. Model 1 = linear growth curve, Model 2 = quadratic growth curve model, Model 3 = quadratic multigroup model, Model 4 = piecewise multigroup model, CFI = comparative fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual, AIC = Akaike information criterion, BIC = Bayesian information criterion. The cutoff criteria for model fit are: CFI ≥ .95, RMSEA < .08, SRMR ≤ .08 (Schreiber, Nora, Stage, Barlow, & King, 2006).

Table II-4, Model 2, and Table II-6, Model 2), we also included a quadratic term in the multigroup models (Model 3). In addition, because the quadratic growth curve shows whether the linear and quadratic growth factors differ between groups, but is not as informative about the timing of differential development, we used piecewise growth curve modeling (Chou et al., 2014) in Model 4 to determine whether the difference in growth stemmed from a certain period of development. Inspecting the estimated and empirical mean curves in Model 3 for both reading and mathematics visually (see Figure II-2), differences in growth between the two groups seemed to exist mainly in the early stages. Thus, we estimated a growth curve with two linear segments, one for the first period (Time 1-Time 2) and one for the time after (Figure II-1B).

Table II-6

Results of Latent Growth Curve Models (LGCMs) for Mathematics Development

	Model 1			Model 2		
	Est	(*SE*)		Est	(*SE*)	
Mean structure						
Intercept	0.03	(0.05)		-0.01	(0.05)	
Linear slope	0.54	(0.02)	***	0.82	(0.05)	***
Quadratic slope				-0.12	(0.02)	***
Covariances						
Intercept – linear slope	0.02	(0.02)		0.03	(0.02)	
Variance structure						
Intercept	0.42	(0.05)	***	0.42	(0.05)	***
Linear slope	0.03	(0.01)	*	0.03	(0.01)	**
Quadratic slope				0.00	(0.00)	

Note. Model 1 = linear growth curve model; Model 2 = quadratic growth curve model; the variance in the quadratic slope was fixed to zero for identification purposes.
*$p < .05$. **$p < .01$. ***$p < .001$.

Reading Achievement

The results in Table II-7 indicate that for children with initially high achievement, the linear and quadratic slope factors in Model 3 differed significantly from the other students. High achievers showed a smaller linear slope ($\Delta_{\text{lin}} = -0.50, p < .001$) and a nonsignificant quadratic growth

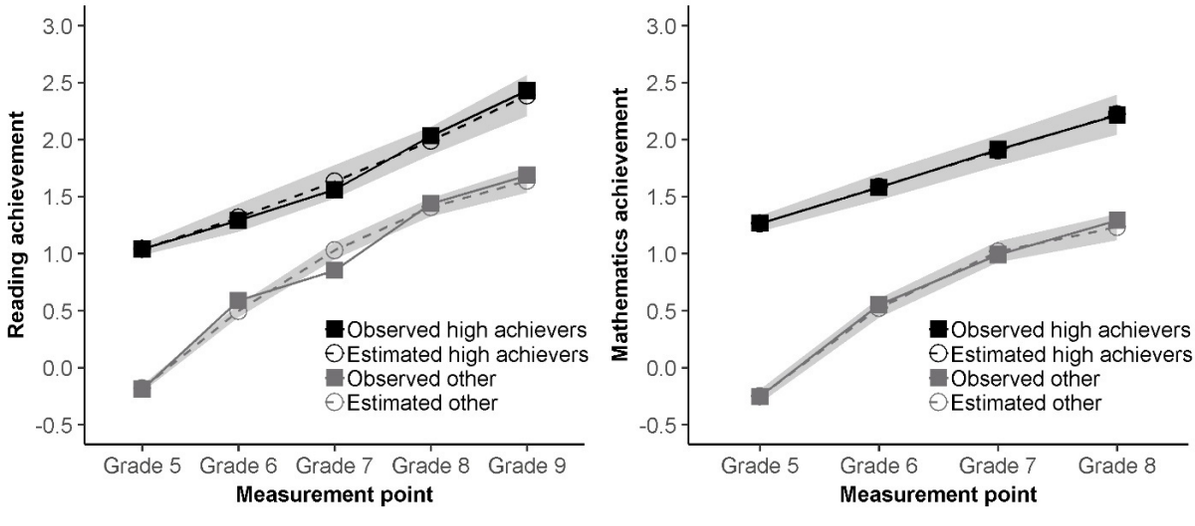


Figure II-2. Sample and estimated means for multigroup models. Solid lines are observed scores, dashed lines are scores estimated in a multigroup quadratic latent growth curve model.

factor that also differed from that of other students ($\Delta_{qua} = 0.10, p < .001$). Thus, the quadratic growth model would provide evidence for a lower growth rate for high achievers and a higher, albeit decelerating, growth rate for the other students and thus for a pattern of compensation.

In order to inspect the timing of differential academic development, the piecewise growth curve was estimated. The first slope was smaller for high-achieving children ($\Delta_{lin1} = -0.53, p < .001$). However, the second slope was not affected by group membership ($\Delta_{lin2} = 0.00, p = .99$). These results suggest that between Grades 5 and 6, compensation was taking place. This compensation was substantial, decreasing the initial achievement gap between both groups by about 42 percent. Afterwards, both groups developed at the same pace.

Table II-8

Results for Nonlinear Multigroup Growth Models in Mathematics Development: Quadratic Model (Model 3) and Piecewise Model (Model 4)

	Quadratic model						Piecewise model					
	High achievers			Other students			High achievers			Other students		
	Est	(*SE*)		Est	(*SE*)		Est	(*SE*)		Est	(*SE*)	
Mean structure												
Intercept	1.26	-0.04	*	-0.25	-0.04	*	1.27	-0.04	*	-0.25	-0.04	*
Linear slope 1	0.32	-0.10	*	0.91	-0.05	*	0.32	-0.08	*	0.82	-0.04	*
Linear slope 2							0.31[a]	-0.06	*	0.41[a]	-0.03	*
Quadratic slope	0.00	-0.04		-0.14	-0.02	*						
Covariances												
Intercept – linear slope 1	0.03	-0.03		0.04	-0.02							
Variance structure												
Intercept	0.06	-0.05		0.22	-0.03	*	0.12	-0.04	*	0.28	-0.03	*
Linear slope 1	0.03	-0.02		0.04	-0.01	*						

Note. The piecewise model's first slope factor identifies the slope between Time 1 and 2; the second slope factor denotes the linear slope between Time 2 and Time 5. The variance in linear slope 1 in the piecewise model and the variance in the quadratic slope in the quadratic model were fixed to zero for identification purposes. Est = estimated parameter; *SE* = standard error. The means of the latent factors differed significantly between the groups, except for the factors marked with the same superscript.
* $p < .01$.

Table II-7

Results for Nonlinear Multigroup Growth Models in Reading Development: Quadratic Model (Model 3) and Piecewise Model (Model 4)

	Model 3						Model 4					
	High achievers			Other students			High achievers			Other students		
	Est	(*SE*)		Est	(*SE*)		Est	(*SE*)		Est	(*SE*)	
Mean structure												
Intercept	1.04	-0.03	*	-0.18	-0.02	*	1.04	-0.03	*	-0.19	-0.02	*
Linear slope 1	0.26	-0.09	*	0.76	-0.04	*	0.26	-0.09	*	0.79	-0.04	*
Linear slope 2							0.38[a]	-0.05	*	0.38[a]	-0.02	*
Quadratic slope	0.02	-0.03		-0.08	-0.01	*						
Covariances												
Intercept-slope 1	0.00	-0.02		0.01	-0.01							
Intercept-slope 2							-0.01	-0.02		0.01	-0.01	
Variance structure												
Intercept	0.09	-0.03	**	0.13	-0.02	***	0.11	-0.03	***	0.13	-0.02	***
Linear slope 1	0.02	0.01		0.03	0.00	*						
Linear slope 2							0.04	0.01	*	0.05	0.01	*

Note. The piecewise model's first slope factor identifies the slope between Time 1 and 2; the second slope factor denotes the linear slope between Time 2 and Time 5. The variance in linear slope 1 in the piecewise model and the variance in the quadratic slope in the quadratic model were fixed to zero for identification purposes. Est = estimated parameter; *SE* = standard error. The means of the latent factors differed significantly between the groups, except for the factors marked with the same superscript.
*p < .05., **p < .01., ***p < .001.

Mathematics Achievement

Results from multigroup models for mathematics (Table II-8) resembled the results for reading. Group membership was related to the shape of the quadratic growth curve such that the difference between the linear growth factors of the two groups was -0.59 (p < .001), and the difference in quadratic growth was 0.14 (p < .001). Thus, high achievers showed a slower growth process but did not show negative acceleration, while the other students exhibited a steeper growth which slowed down over time (Figure II-2). The piecewise Model 4 showed that the first slope in mathematics skill development was lower for high achievers ($\Delta_{\text{lin1}} = -0.49, p < .001$), decreasing the original achievement gap by about 31 percent, but the groups did not differ in the second slope ($\Delta_{\text{lin2}} = -0.09, p = .16$). The differences between the two groups were also confirmed by an alternative model in which we estimated direct paths from a dummy group

membership factor to the growth factors rather than using a multigroup approach (details can be found in Supplement C).

Robustness checks

To determine in how far results are a product of choices made during data analysis, we conducted a series of further analyses in which we systematically changed factors that may have influenced the results, namely the methods used when scaling the data, the method for dealing with missing data, and the cut-off criterion used for defining high achievers. We thereby followed the idea of a multiverse analysis proposed by Steegen, Tuerlinckx, Gelman and Vanpaemel (2016) as well as the specification curve approach of Simonsohn, Simmons and Nelson (2015). Our results proved to be robust, with all effects pointing in the same direction. More specifically, the standardized slope differences between both groups had a mean of $\bar{d}_{\text{lin1}} = 1.22$ across all analysis models in reading and a mean of $\bar{d}_{\text{lin1}} = 0.86$ across all models in mathematics. The details of these analyses and results are in Supplement II-B.

Discussion

In this study, our goal was to examine the differential academic development of high-achieving students in secondary school in both reading and mathematics. Based on previous research, competitive hypotheses were formulated predicting (H1.a) that high achievers would extend their lead with respect to their achievement (cumulative advantage; Matthew effect), (H1.b) that previously lower achieving classmates would catch up over time (compensation), or (H1.c) that the two groups would develop at the same pace and in parallel.

Overall, our results did not support the Matthew effect hypothesis in high-track secondary schools in either reading or mathematics (H1a. rejected). Our results rather supported the compensation hypothesis for both reading and mathematics because initially high-achieving students showed less growth than other students (H1.b sustained). The results are aligned with earlier work that found Matthew effects in elementary education but not for an extended period of time after the children started school (Baumert et al., 2012). The effect sizes we found in our sample were quite high compared with other studies. A reason for this might be the relative homogeneity of our sample

The absence of Matthew effect patterns in our study could have a variety of causes. On the one hand, high-achieving students might not have been challenged enough and might not have found sufficient opportunities to develop their skills in school. Although high-track secondary schools are supposed to serve the highest achieving students in the educational system, teachers do not seem to succeed in enabling these students to progress more rapidly. Different studies have shown that on average, teachers differentiate very little and do not seem to orientate their teaching toward

the most able students but rather to an imaginary average student (Archambault et al., 1993; Jennek et al., 2018). The supposition that high-track secondary schools do not optimally support high achievers has been supported by research showing similar patterns of academic development in high achievers who make an early transition into Gymnasium compared with comparable students who stay in primary school for 2 more years, a choice that is possible in certain German states (Baumert et al., 2009). On the other hand, for those high-track students with relatively lower achievement, the positive effects of transitioning to the high-track schools seem to have outweighed the negative effects of a lower achievement rank so that they were not at a disadvantage as might be expected.

It is possible that teachers have played an active part in aiming for achievement homogeneity in their classes: Depending on their conception of justness, teachers allocate opportunities to individual students according to the principles of need, justness, or equality (Schwippert & Walker, 2003). If teachers are motivated by the need principle, they should allocate more attention and resources to students with lower achievement to help ensure that they do not fall further behind. Accordingly, Nurmi, Viljaranta, Tolvanen, & Aunola (2012) as well as Kiuru et al. (2015) have shown for primary teachers that they directed more active instruction to lower performing students than to high performers. For specific tracks in secondary school, such research is missing. However, it is not only that the individual teachers in class might have distributed their efforts unevenly. It might also be the case that some students with lower achievement in high-track schools received additional remediation or out-of-school tutoring, which helped them keep up or even improve substantially. For German primary schools, it has been shown that schools are more likely to provide remedial tutoring for students with difficulties than they are to provide extracurricular enrichment opportunities for high-achieving students (Neuendorf, Kuhl, & Jansen, 2017). Future research is needed to investigate whether variations in the use of differentiated and adaptive instruction or enrichment opportunities in school make a difference in the achievement trajectories of high-achieving students. Also, further research could investigate whether the abovementioned speculations hold and teachers are deliberately or unwittingly less likely to promote high achievers than other students, especially in the early grades of secondary school.

The results in reading and mathematics are in agreement concerning the lack of Matthew effects within the highest secondary school track (H3 rejected). This could either mean that the same processes are at play in both domains, but the compensation effects could also be stemming from different causes leading to the same result in both subjects. Reading practice, for example, is less confined to the school context: Secondary school students need reading competencies in their everyday life. Consequently, the additional time students read outside of school or in other school

subjects might help them catch up more quickly. In mathematics, students having problems with the lesson content might be more easily identified as tasks usually have a well-defined solution and mistakes can be easily detected. Subsequently, teachers can give special attention to these students. Both reading and mathematics skills might have in common that students differ in their onset of development and developmental trajectories similarly follow a pattern of slowing growth across the school career (Bloom, Hill, Black, & Lipsey, 2008), resulting in catching-up effects.

Our analyses confirmed that growth trajectories as well as differences therein followed a non-linear shape (H2 sustained). Differences in growth between the two achievement groups seemed to occur mainly in the first period of measurement, that is, from Grade 5 to Grade 6. On the one hand, it is plausible that, especially in the first year after the transition, students adjust to the new demands placed on them, and teachers have sufficiently homogenized the students' achievement levels so that further learning can proceed at a more even pace for all. On the other hand, compensation between the first two measurement occasions is a pattern that we would expect given the effect of regression to the mean. Those scoring highest on a first test will, for statistical reasons, score lower on average when measured a second time. Therefore, the compensation effects we found could also be statistical artefacts. This is a fundamental problem when analyzing the development of extreme groups over time (Campbell & Kenny, 2003). One possibility suggested in the literature is the usage of residualized change scores as a supplement to regular growth analysis. Thus, in an additional analysis, which is included in Supplement D, we calculated residualized change scores, but we cannot rule out that results are partly caused by regression to the mean.

Limitations

Our study has some limitations. First, there was a great deal of attrition between Grades 7 and 8. This dropout was mainly due to the requirement of a new informed consent form after Grade 7 (Homuth et al., 2017). We used the FIML procedure, which actually requires missing values to occur "at random" to deal with the missing data (Enders, 2010; see Section 2.4). However, Homuth et al. (2017) showed that dropout was selective in that, among other reasons, students with less educated parents and a lower ISEI score were more likely to terminate their attendance. Enders (2011) recommended that a sensitivity analysis be performed by estimating a similar model using multiple competing missing data treatment methods. In a multiverse follow-up analysis, we therefore additionally fit the selection model to the data (Enders, 2011). This led to essentially the same results, which we have included in Supplement B.

Further, the model fit in reading was not satisfactory (see Table II-5). As the inspection of the observed curves showed, the curve did not closely follow a linear or quadratic shape. We accepted imposing a less-than-optimal model to the data in the interest of answering our substantive question, especially in comparing the model results between reading and mathematics. However, with more points of measurement per year, it might be easier to describe the shape of the curve in reading because more measurement occasions allow for greater flexibility in model building.

When conducting research on extreme groups, the interpretation of results rests on the ability of the instrument to properly differentiate between students at the ends of the distribution. Regrettably, the test instruments used in the BiKS-study differentiate better at the lower ends of the distribution. Particularly the reading test suffers from ceiling effects at T2 and T3. Of course, this is concerning, as the effect of compensation could be caused by those high achievers not being able to demonstrate their competencies fully. Future research should strive to develop and apply tests which are able to properly assess achievement at all levels. Adaptive testing would be a possibility to increase test information and reduce measurement error without having the problem of too long tests or motivational problems in students for whom the test is too difficult.

Another limitation of our study is the small number of linking items. According to Angoff (1984), as a rule of thumb, no less than 20 items should be used. Whereas a larger number of linking items might reduce the standard error of linking, there are different views on the criteria for including or excluding linking items (Robitzsch, Dörfler, Pfost, & Artelt, 2011; Taherbhai & Seo, 2013). We checked whether different recommended strategies (using all common items as linking items; discarding items showing DIF from the linking pool but leaving them in the tests; discarding items showing DIF from the tests altogether) would change the results (see Supplement B). We found an effect on the magnitude of effects in reading, but not on their direction. Moreover, the rank-order stability in freely estimated anchor items was between .80 and .96 in reading and between .94 and .97 in mathematics. We therefore concluded that the anchor items were sufficiently consistent across waves. However, future research could substantiate the findings if longer tests could be used.

A point that could be criticized in our study is the use of a cut-off criterion to categorize students into the group of high achievers. For one thing, achievement is obviously a continuum, and whether there is a threshold at which a qualitative difference in learning and development processes between those above and below an arbitrary cut-off can be observed seems disputable. We paid attention to this fact by estimating Models 1 and 2 for the total sample. Here, the nonsignificant correlation between the intercept and slope would imply that parallel development

had taken place. Models 3 and 4, in which students were grouped according to our cut-off criterion, indicated compensatory development. These different results seem to legitimize the use of categories for analytical purposes when inferences should be drawn about high achievers. However, the question of whether it is possible to find subgroups of students that follow distinct developmental patterns and whether these subgroups are related to different initial achievement levels should be the subject of future research, for example, by using latent class growth analyses (Jung & Wickrama, 2008).

Conclusion

Because students come to school with different skills, it is a challenge for teachers to meet the needs of children who differ substantially in prior achievement. The transition to secondary school in Germany signifies a major change in the educational context—in peers, in teachers, and in teaching practices—that could have a differential impact on patterns of development. It was our aim to clarify whether high-achieving students show different curves for reading and/or mathematics development than other students, whether these other students catch up during secondary school, or whether the two groups show similar rates of growth. Our results favor the compensation hypothesis for both domains. Thus, we found no support for the notion of Matthew effects in high-track secondary school in reading or in mathematics.

In how far these results can be generalized remains an open question. While some explanatory approaches to the results are universal in nature (for example the lagged non-linear skill development), others might be confined to this specific context of Germany and to these specific competencies under study (for example the employment of certain teaching strategies and the perceived role of a teacher). Thus, replicating our findings within other contexts could help in determining the extent to which these findings are generalizable.

Whereas the teachers of this sample of students seem to be successful in fostering academic progress in all students similarly, it might be expected that students with high initial competencies can learn at a faster rate when challenged adequately. Therefore, the promotion of high achievers in lower secondary school might profit from research on the effectiveness of applying differentiating strategies in secondary school.

References

Aarnoutse, C., & van Leeuwe, J. (2000). Development of poor and better readers during the elementary school. *Educational Research and Evaluation*, *6*(3), 251–278. https://doi.org/10.1076/1380-3611(200009)6:3;1-A;FT251

Aarnoutse, C., van Leeuwe, J., Voeten, M., & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing: An Interdisciplinary Journal*, *14*, 61–89.

Archambault, F. X., Westberg, K. L., Brown, S. W., Hallmark, B. W., Zhang, W., & Emmons, C. L. (1993). Classroom practices used with gifted third and fourth grade students. *Journal for the Education of the Gifted*, *16*(2), 103–119. https://doi.org/10.1177/016235329301600203

[dataset] Artelt, C., Blossfeld, H.-P., Faust, G., Roßbach, H.-G., & Weinert, S. (2013). *Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (BiKS-8- 14)* [Educational processes, competence development and selection decisions in preschool and school age]. Institut zur Qualitätsentwicklung im Bildungswesen. Datset, https://doi.org/10.5159/IQB_BIKS_8_14_v2.

Bannister, N. A. (2016). Breaking the spell of differentiated instruction through equity pedagogy and teacher community. *Cultural Studies of Science Education*, *11*(2), 335–347. https://doi.org/10.1007/s11422-016-9766-0

Bast, J., & Reitsma, P. (1997). Mathew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research*, *32*(2), 135–167. https://doi.org/10.1207/s15327906mbr3202_3

Baumert, J., Becker, M., Neumann, M., & Nikolova, R. (2009). Frühübergang in ein grundständiges Gymnasium – Übergang in ein privilegiertes Entwicklungsmilieu? [Early transition into the academic track of secondary schooling – Transfer into a privileged learning environment?]. *Zeitschrift Für Erziehungswissenschaft*, *12*(2), 189–215. https://doi.org/10.1007/s11618-009-0072-4

Baumert, J., Lehmann, R., Lehrke, M., Clausen, M., Hosenfeld, I., Neubrand, J., . . . Günther, W. (1998). *Testaufgaben Mathematik TIMSS 7./8. Klasse (Population 2)*. Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Maaz, K., Stanat, P., Watermann, R. (2009). Schulkomposition oder Institution –was zählt? Schulstrukturen und die Entstehung schulformspezifischer Entwicklungsverläufe [Compositional or institutional factors—What counts at school? School structures and the Emergence of track-specific developmental trajectories]. *Die Deutsche Schule, 101*(1), 33-46.

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? *Child Development*, *83*(4), 1347–1367. https://doi.org/10.1111/j.1467-8624.2012.01779.x

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the emergence of a differencial learning and development environment]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90082-7_4

Becker, M., Neumann, M., Tetzner, J., Böse, S., Knoppick, H., Maaz, K., . . . Lehmann, R. (2014). Is early ability grouping good for high-achieving students' psychosocial development? Effects

of the transition into academically selective schools. *Journal of Educational Psychology, 106*(2), 555–568. https://doi.org/10.1037/a0035425

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions* (MDRC Working Papers on Research Methodology). Retrieved from MDRC website: https://www.mdrc.org/sites/default/files/full_473.pdf

Brighton, C. M., Hertberg, H. L., Moon, T. R., Tomlinson, C. A., & Callahan, C. M. (2005). *The feasibility of high-end learning in a diverse middle school* (RM05210). Storrs, Conneticut: National Research Center on the Gifted and Talented.

Campbell, D. T., & Kenny, D. A. (2003). A primer on regression artifacts: Donald T. Campbell, David A. Kenny (Reprinted.). Methodology in the social sciences. New York: Guilford Press.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation*, *49*, 30–41. https://doi.org/10.1016/j.stueduc.2016.03.005

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. *The American Psychologist*, *60*(2), 149–160. https://doi.org/10.1037/0003-066X.60.2.149

Chou, C.-P., Yang, D., Pentz, M. A., & Hser, Y.-I. (2004). Piecewise growth curve modeling approach for longitudinal prevention study. *Computational Statistics & Data Analysis*, *46*(2), 213–225. https://doi.org/10.1016/S0167-9473(03)00149-X

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record*, *115*(6), 1–29.

Cunningham, Anne E.; Stanovich, Keith E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Dar, Y., & Resh, N. (1986). Classroom intellectual composition and academic achievement. *American Educational Research Journal*, *23*(3), 357–374. https://doi.org/10.3102/00028312023003357

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

DFG-Forschergruppe „Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter“ (BiKS). (n.d.). *Codebuch zur Kompetenzerhebung Welle 4: BiKS-8-14 Sekundarstufe* [Codebook for Competence Assessment Wave 4: BiKS-8-14 Secondary School]. Retrieved from https://www.iqb.hu-berlin.de/fdz/studies/BiKS_8-14/Kompetenzerhebun_3.pdf

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, *32*(1), 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Dockx, J., Fraine, B. de, & Vandecandelaere, M. (2018). Does the track matter? A comparison of students' achievement in different tracks. *Journal of Educational Psychology.* Advance online publication. https://doi.org/10.1037/edu0000305

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. *Multivariate applications*. Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Enders, C. K. (2010). Applied missing data analysis. Methodology in the social sciences. New York: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16. https://doi.org/10.1037/a0022640

Fischer, G. H. (1995). Linear logistic models for change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 157–180). New York: Springer.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*(1), 3–17. https://doi.org/10.1037/0022-0663.88.1.3

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*, *104*(4), 1094–1108. https://doi.org/10.1037/a0028389

Gagné, F. (2009). Building gifts into talents: Detailed overview of the DMGT 2.0. In B. MacFarlane & T. Stambaugh (Eds.), *Leading change in gifted education: The festschrift of Dr. Joyce VanTassel-Baska* (pp. 61–80). Waco, TX: Prufrock Press.

Ganzeboom, H. B., Graaf, P. M. de, & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Götz, L., Lingel, K., & Schneider, W. (2013). *DEMAT 6+: Deutscher Mathematiktest für sechste Klassen* [German Mathematics Test for Sixth Classes]: Hogrefe.

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

Heimendinger, J., & Laird, N. (1983). Growth changes: Measuring the effect of an intervention. *Evaluation Review*, *7*(1), 80–95. https://doi.org/10.1177/0193841X8300700105

Homuth, C., Schmitt, M., Lorenz, C., & Mann, D. (2017). Warum ein erneutes Genehmigungsverfahren im laufenden Längsschnitt weitreichende Folgen für die Datenqualität hat [Why does a new consent procedure during an ongoing longitudinal study influence data quality]. *Journal for Educational Research Online*, *9*(1), 7–31. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/viewFile/731/302

Jennek, J., Gronostaj, A., & Vock, M. (2018). Wie Lehrkräfte im Englischunterricht differenzieren. Eine Re-Analyse der DESI-Videos [Differentiation among German teachers of English: Reanalyzing the DESI Study videos]. *Unterrichtswissenschaft.* Advance online publication. https://doi.org/10.1007/s42010-018-0027-7

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development*, *74*(3), 834–850. https://doi.org/10.1111/1467-8624.00571

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Kiss, D. (2013). The impact of peer achievement and peer heterogeneity on own achievement growth: Evidence from school transitions. *Economics of Education Review*, *37*, 58–65. https://doi.org/10.1016/j.econedurev.2013.08.002

Kiuru, N., Nurmi, J.-E., Leskinen, E., Torppa, M., Poikkeus, A.-M., Lerkkanen, M.-K., & Niemi, P. (2015). Elementary school teachers adapt their instructional support according to students' academic skills: A variable and person-oriented approach. *International Journal of Behavioral Development*, *39*(5), 391–401. https://doi.org/10.1177/0165025415575764

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer New York.

Kühn, P., Sachse, S., & Suchodoletz, W. von. (2015). Sprachentwicklungsverzögerung: Was wird aus Late Bloomern? [Language delay: What is the prognosis of late bloomers?]. *Klinische Padiatrie*, *227*(4), 213–218. https://doi.org/10.1055/s-0035-1547310

Kulik, C.-L. C., & Kulik, J. A. (1982). Effects of Ability Grouping on Secondary School Students: A Meta-analysis of Evaluation Findings. *American Educational Research Journal*, *19*(3), 415–428. https://doi.org/10.3102/00028312019003415

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., . . . Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler [Quality of mathematics instruction across school types: Findings from PISA 2003]. *Zeitschrift Für Erziehungswissenschaft*, *8*(4), 502–520. https://doi.org/10.1007/s11618-005-0156-8

Li, A. K. F., & Adamson, G. (1992). Gifted Secondary Students' Preferred Learning Style: Cooperative, Competitive, or Individualistic? *Journal for the Education of the Gifted*, *16*(1), 46–54. https://doi.org/10.1177/016235329201600106

Linchevski, L., & Kutscher, B. (1998). Tell me with whom you're learning, and I'll tell you how much you've learned: Mixed-ability versus same-ability grouping in mathematics. *Journal for Research in Mathematics Education*, 533–554. https://doi.org/10.2307/749732

Lorenz, C., Schmitt, M., Lehrl, S., Mudiappa, M., & Rossbach, H.-G. (2013). The Bamberg BiKS research group. In M. Pfost, C. Artelt, & S. Weinert (Eds.), *Schriften aus der Fakultät Humanwissenschaften der Otto-Friedrich-Universität Bamberg - 14. The Development of Reading Literacy from Early Childhood to Adolescence. Empirical Findings from the Bamberg BiKS Longitudinal Studies* (pp. 15–34). Bamberg: University of Bamberg Press.

Lu, Y. (2016). Modeling Math Growth Trajectory—An Application of Conventional Growth Curve Model and Growth Mixture Model to ECLS K-5 Data. *Journal of Educational Issues*, *2*(1), 166. https://doi.org/10.5296/jei.v2i1.9197

Ma, X. (2005). A longitudinal assessment of early acceleration of students in mathematics on growth in mathematics achievement☆. *Developmental Review*, *25*(1), 104–131. https://doi.org/10.1016/j.dr.2004.08.010

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*(2), 99–106. https://doi.org/10.1111/j.1750-8606.2008.00048.x

Magrini, A. (2016). dlsem: Distributed-Lag Structural Equation Modelling. Retrieved from https://CRAN.R-project.org/package=dlsem

Marx, H., & Opitz-Karig, U. (2005). *DEMAT5+: Deutscher Mathematiktest für fünfte Klassen* [German Mathematics Test for Fifth Classes]. Leipzig, Germany.

McCoach, D. Betsy; Siegle, Del (2007): What Predicts Teachers' Attitudes Toward the Gifted? In: *Gifted Child Quarterly 51*(3), S. 246–254. DOI: 10.1177/0016986207302719.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, *44*(5), 472–488. https://doi.org/10.1177/0022219411414010

Morgan, P. L., & Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional Children*, *73*(2), 165–183.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of learning disabilities*, *51*(6), 523–539.

Neuendorf, C., Kuhl, P., & Jansen, M. (2017). Leistungsstarke Schülerinnen und Schüler in Deutschland [High-achieving students in Germany]. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich, & N. Haag (Eds.), *IQB-Bildungstrend 2016: Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (pp. 317–334). Münster: Waxmann.

Nurmi, J.-E., Viljaranta, J., Tolvanen, A., & Aunola, K. (2012). Teachers adapt their instruction according to students' academic performance. *Educational Psychology*, *32*(5), 571–588. https://doi.org/10.1080/01443410.2012.675645

OECD. (2016). PISA 2015 results (Volume I): Excellence and equity in education. PISA. Paris: OECD Publishing.

Pekrun, R., Götz, T., Jullien, S., Zirngibl, A., vom Hofe, R., & Blum, W. (2003). *Skalenhandbuch PALMA: 2. Messzeitpunkt (6. Klassenstufe)*. München.

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research*, *84*(2), 203–244. https://doi.org/10.3102/0034654313509492

Pfost, M., & Artelt, C. (2013). Reading literacy development in secondary school and the effect of differential institutional learning environments. In M. Pfost, C. Artelt, & S. Weinert (Eds.), *Schriften aus der Fakultät Humanwissenschaften der Otto-Friedrich-Universität Bamberg - 14. The Development of Reading Literacy from Early Childhood to Adolescence. Empirical Findings from the Bamberg BiKS Longitudinal Studies* (pp. 229–277). Bamberg: University of Bamberg Press.

Pfost, M., Dörfler, T., & Artelt, C. (2010). Der Zusammenhang zwischen außerschulischem Lesen und Lesekompetenz [The relation between extra-curricular reading behavior and reading competence: Results from a longitudinal study at the transition from primary to secondary school]. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *42*(3), 167–176. https://doi.org/10.1026/0049-8637/a000017

Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample: an empirical examination of the Matthew effect model. *Journal of Research in Reading*, *35*(4), 411–426. https://doi.org/10.1111/j.1467-9817.2010.01478.x

Pfost, M., Dörfler, T., & Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learning and Individual Differences*, *26*, 89–102. https://doi.org/10.1016/j.lindif.2013.04.008

Pfost, M., Karing, C., Lorenz, C., & Artelt, C. (2010). Schereneffekte im ein- und mehrgliedrigen Schulsystem [Fan spread effects in a tracked and a nontracked school system]. *Zeitschrift Für Pädagogische Psychologie*, *24*(3-4), 259–272.

Phillips, N., & Lindsay, G. (2006). Motivation in gifted students. *High Ability Studies*, *17*(1), 57–73. https://doi.org/10.1080/13598130600947119

Plucker, J. A., & Callahan, C. M. (2014). Research on giftedness and gifted education: Status of the field and considerations for the future. *Exceptional Children*, *80*(4), 390–406. https://doi.org/10.1177/0014402914527244

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In Search of Matthew Effects in Reading. *Journal of Learning Disabilities*, *49*(5), 499–514. https://doi.org/10.1177/0022219414559974

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rambo-Hernandez, K. E., & McCoach, D. B. (2014). High-Achieving and Average Students' Reading Growth: Contrasting School and Summer Trajectories. *The Journal of Educational Research*, *108*(2), 112–129. https://doi.org/10.1080/00220671.2013.850398

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen [Relevance of item selection and model selection for assessing the development of competencies: The development in reading competence in primary school students]. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *43*(4), 213–227. https://doi.org/10.1026/0049-8637/a000052

Scammacca, N., Fall, A.-M., Capin, P., Roberts, G., & Swanson, E. (2019). Examining factors affecting reading and math growth and achievement gaps in grades 1–5: A cohort-sequential longitudinal approach. *Journal of Educational Psychology.* Advance online publication. https://doi.org/10.1037/edu0000400

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schwippert, K., & Walker, M. (2003). Homogenous and high performing classes: The case of optimal classes. *Studies in Educational Evaluation*, *29*(2), 109–128. https://doi.org/10.1016/S0191-491X(03)00018-X

Shaywitz, B. A., Holford, T. R., Holahan, J. M., Fletcher, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A matthew effect for IQ but not for reading: Results from a longitudinal study. *Reading Research Quarterly*, *30*(4), 894–906.

Shin, T., Davison, M. L., Long, J. D., Chan, C.-K., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences*, *23*, 92–100. https://doi.org/10.1016/j.lindif.2012.10.002

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.2694998

Skibbe, L. E., Grimm, K. J., Stanton-Chapman, T. L., Justice, L. M., Pence, K. L., & Bowles, R. P. (2008). Reading trajectories of children with language difficulties from preschool through fifth grade. *Language, Speech, and Hearing Services in Schools*, *39*, 475–486.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*(3), 471–499. https://doi.org/10.2307/1170761

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology*, *76*(3), 347–376. https://doi.org/10.1037/0022-0663.76.3.347

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *XXI*(4), 360–406.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Taherbhai, H., & Seo, D. (2013). The philosophical aspects of IRT equating: Modeling drift to evaluate cohort growth in large scale assessments. *Educational Measurement: Issues and Practice*, *32*(1), 2–14.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., . . . Reynolds, T. (2003). Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A Review of Literature. *Journal for the Education of the Gifted*, 27(2-3), 119–145. https://doi.org/10.1177/016235320302700203

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: how students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology*, *101*(4), 853–866. https://doi.org/10.1037/a0016306

Udvari, S. J., & Schneider, B. H. (2000). Competition and the adjustment of gifted children: A matter of motivation. *Roeper Review*, *22*(4), 212–216. https://doi.org/10.1080/02783190009554040

Von Davier, M., & von Davier, A. A. (2004). *A unified approach to IRT scale linking and scale transformation*. Princeton, New Jersey: ETS.

Walberg, H. J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*(3), 359–373.

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

Weirich, S., & Hecht, M. (2018). eatModel. Retrieved from https://R-Forge.R-project.org/projects/eat/

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219-240, 269-281). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Wu, M. L., Adams, R. J., Wilson, & Haldane, S. A. (2007). ACER conquest version 2.0: Camberwell, Victoria, Australia: ACER Press, Australian Council for Educational Research.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Zuckerman, H. (2011). The Matthew Effect Writ Large and Larger: A Study in Sociological Semantics. In Y. Elkana, A. Szigeti, & G. Lissauer (Eds.), *Robert K. Merton and the Future of Sociology. Concepts and the Social Order* (1st ed., pp. 121–164). Central European University Press.

Supplemental Material

Supplement A: The German secondary school system

After primary school, German students enter a tracked secondary school system. Because the German federal states (rather than the national government) are in charge of regulating the school systems, there are large differences between the federal states with regard to school tracks and types. For example, in some states, students enter the secondary school system after 4 years of primary school, whereas in other states, 6 years of primary school are mandatory. Also, the traditional three-tier structure with *Gymnasium* as the primary academic, university-bound track, *Realschule* as the intermediate track, and *Hauptschule* as the lowest vocationally oriented track is still dominant in some federal states, whereas in most states, efforts are being made to achieve a stronger integration of school types. Some states differentiate between only the academic track (*Gymnasium*) and one other secondary school type (e.g., *Sekundarschule*) in which within-school tracking is employed. In Bavaria and Hesse, all students enter the secondary school system after Grade 4. In the first grade of primary school, the majority of students are 6 years old, so they move on to the secondary school system on average at age 10. The decision about which type of secondary school a student will enter is based on teachers' recommendations and parents' preferences. In Bavaria and Hesse, the three-tier system still exists, with a small number of additional school types that employ within-school tracking or have a specific vocational focus (*Mittelstufenschule*, *Gesamtschule*, *Wirtschaftsschule*). A central characteristic of the secondary school system in Germany that is present in all federal states is the differentiation between *Gymnasium* and the other more vocationally oriented school types (also referred to as the nonacademic track).

Supplement B: Multiverse Analysis and Specification Curve

In analyzing data, a number of choices have to be made regarding data preparation and model building. These choices may have an impact on the results of a study and might be a cause for limited reproducibility. To increase transparency and robustness of results, Steegen, Tuerlinckx, Gelman and Vanpaemel (2016) proposed so-called multiverse analysis in which the choices made during the data analytical phase of a study are systematically altered and all reasonable combinations of specifications are made and the resulting effects are compared across all specifications.

Our analyses were characterized by a number of choices made during the scaling of the reading and mathematics tests and in our substantive analyses. We have listed all of these choices in which we think reasonable alternatives could have been chosen (see Table II-B1). These possible choices will first be explained in the following, before turning to the results of our analysis.

Table II-B1

Specification Factors in the Current Study

Specification factor	Possible alternatives
Sample used for scaling the test	Full sample **Analysis sample only**
Method used for anchoring the scales of subsequent tests	Use all anchor items **Use only items without item parameter drift as anchors** Exclude items displaying item parameter drift from the test
Method used for linking	**Fixed item parameter calibration (FCIP) method** Linking according to Haberman (2009)
Method used for dealing with missing values	**Full Information Maximum Likelihood (FIML) method** Selection model
Cut-off criterion for defining high achievers	Top 10% **1 SD above the sample mean** Top 25% Top 50%

Note. The specifications used in this paper are printed in bold face.

When calibrating the test items, information on the proportion of examinees answering an item correctly is used to estimate the difficulty of each test item. On the one hand, the size of the sample and the dispersion of ability in the sample used for calibrating the test is important for obtaining estimates of a higher precision (Hambleton & Cook, 1983; Stocking, 1990). Therefore, using the full sample for item parameter calibration might be a reasonable choice. On the other hand, we wanted the constructed scale to capture the development of high-track students and given their different learning environment at these schools, they might be a non-comparable subpopulation. This view is further supported by the fact that additional items specifically designed for high track and other items specifically lower tracks were used in the study. It might therefore be expected that their developmental path differs so much from others that the longitudinal linking precision suffers from the diverse sample.

When connecting the scales of the tests longitudinally, two aspects need to be taken into account: the choice of anchoring items and the statistical method used for linking. Anchor items are items which are administered in successive waves of measurement and their item parameters are used to link the scales of the tests used in successive waves. The alignment of scales is a precondition for making assertions about longitudinal competence development and comparing achievement scores across time (for a discussion, see Protopapas et al., 2016). Thus, it has to be ensured that items in successive waves have the same relative difficulty in different administrations. Changes in item difficulty between administrations is called item parameter drift and can be detrimental to obtaining comparable ability estimates (Rupp & Zumbo, 2006). Therefore, these items may be excluded from the linking pool or from the test altogether. However, some authors also argue for including items displaying item drift in certain cases (Robitzsch, Dörfler, Pfost & Artelt, 2011; Taherbhai & Seo, 2013).

When appropriate anchors have been chosen, different methods for conducting the linking are available. We have used the fixed item parameter calibration method, in which basically, the difficulty parameters of anchor items are fixed to the value obtained in the prior wave. Another option would be the linking method according to Haberman (2009) in which all waves are calibrated separately and then their scales are aligned subsequently by the means of the common (anchor) items.

After the tests have been calibrated and linked and the person ability scores have been computed, the question is how to proceed with missing values. Most statistical models assume that values are missing at random (MAR) or completely at random (MCAR). However, this assumption is not always tenable. Enders (2011) has suggested the use of multiple methods of missing data treatment

to evaluate whether results are invariant across the chosen models. Here, we use the FIML method implemented in MPlus (which assumes the MAR mechanism) and the selection model described by Enders (2011) as a model used with missing not at random (MNAR) data. In the selection model, missing data indicators are specified for each measurement and these are regressed on the outcome at the previous and the concurrent measurement occasions.

Finally, if students are grouped into high achievers and other students, the level at which to partition the achievement scale can be a point of discussion. Our choice of a cut-off value at 1 SD above the sample mean was motivated by identifying a population that is comparable to high-achievers as defined by the highest levels in recent large scale assessments. However, in our multiverse analysis we included additional cut-off values at the upper half, the upper quartile and the top 10 % students of the ability distribution to find out how much the result varies depending on the grouping of students into categories.

The combination of all alternative choices leads up to 96 models to be analyzed for each reading and mathematics. We have estimated the piecewise model presented in Section 3.3 (Figure II-1B). We then calculated an effect size for the difference between the two group's slopes for each model as follows:

$$d_{Lin1} = \frac{Lin1_{HA} - Lin1_{other}}{\sqrt{\frac{((n_{HA} - 1) \times {sd_{HA}}^2) + ((n_{other} - 1) \times {sd_{other}}^2)}{N - 2}}}$$

This formula is based on Feingold (2009), who recommends to use the raw score SD to calculate effect sizes for growth model analyses. We have pooled the SD between the groups and corrected for different group sizes according to Hedges & Olkin (1985).

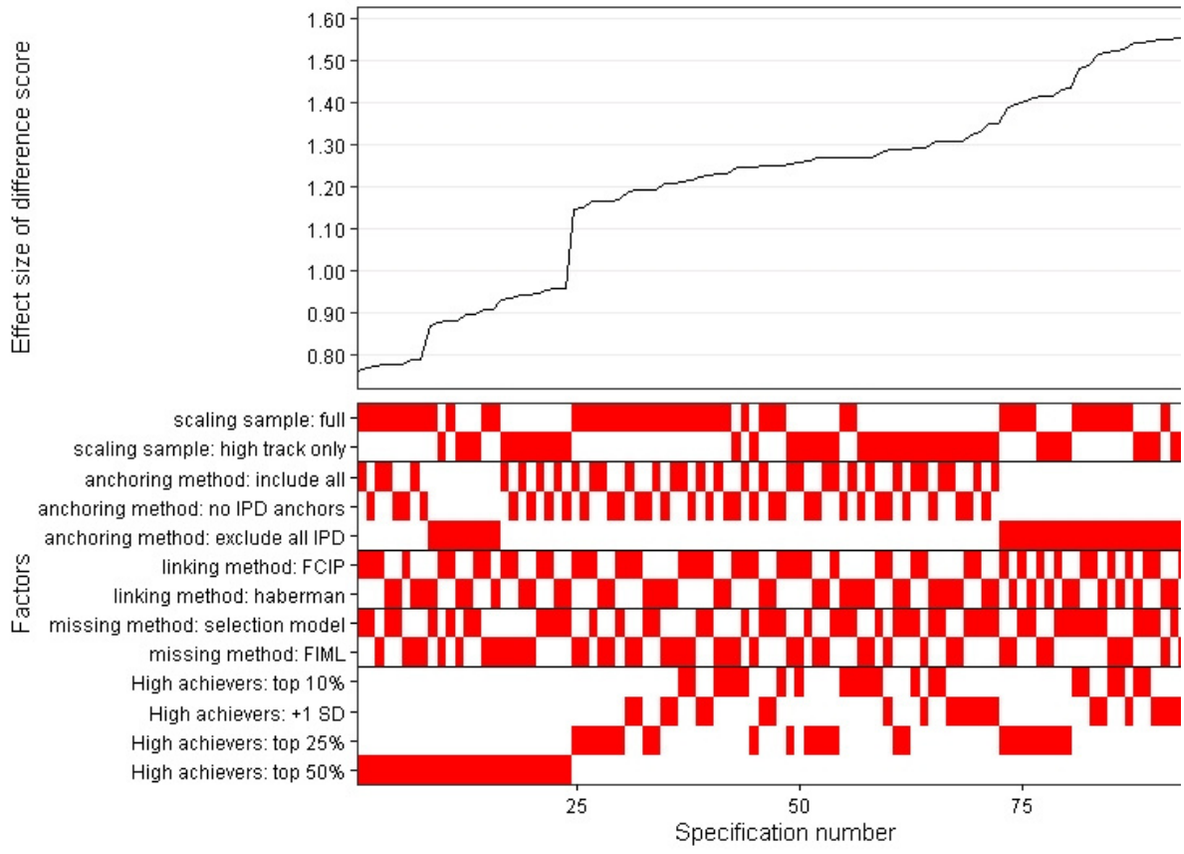


Figure II-B1. Descriptive specification curve for reading. The vertical columns represent the specifications with the chosen value on each of the factors indicated by a red color in the bottom panel. The corresponding effect size estimate for each specification is depicted on the top panel. A total of 96 specifications were estimated.

The results are depicted in a specification plot (Simonsohn et al., 2015). Each column in the plot corresponds to one specification. For reading, effect sizes of between group differences in rates of growth, i.e. slopes range between $d = 0.76$ and $d = 1.59$. From Figure II-B1, it can be seen that the greatest leap in the curve is caused by the definition of high achievers. Partitioning the sample at the mean leads to the smallest differences between the groups' slopes which are $d = .96$ at most. In selecting the top 25 %, effect sizes start from $d = 1.15$. When selecting a more extreme sample, the effect sizes become larger, but there is a large overlap between the choices. Interestingly, when items that show item parameter drift are excluded from the sample of items, the effect sizes are highest. This shows that these analyses are very sensitive to changes in the number of items. The scaling sample affects results to a certain extent: In general, effect sizes are smaller when the full sample was used in calibrating the test. However, the method used to account for missingness and the model used for linking seemed to have no major impact on the findings.

The specification curve for mathematics can be seen in Figure II-B2. Here, effect sizes of between group differences in mathematics growth were between $d = 0.54$ and $d = 1.14$. The factors

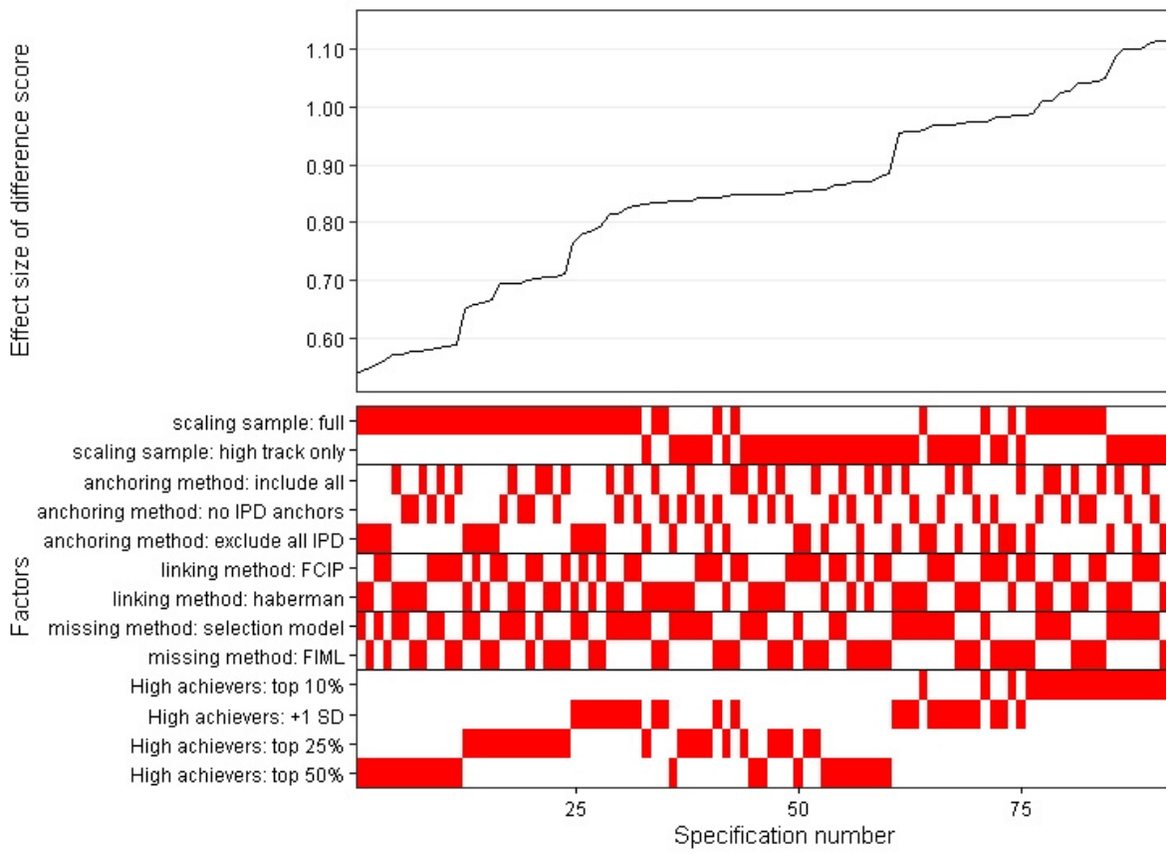


Figure II-B2. Descriptive specification curve for mathematics. The vertical columns represent the specifications with the chosen value on each of the factors indicated by a red color in the bottom panel. The corresponding effect size estimate for each specification is depicted on the top panel. A total of 96 specifications were estimated.

affecting the size of effects differed somewhat from reading. First, the analyses were very sensitive to the scaling sample used, with the full sample resulting in smaller effects than using the high track students only. The second important point was the selection of high achievers. It can be seen that the points at which there is a sudden increase in effect sizes are at the borders between the grouping factor levels. The way we proceeded with anchoring items showing item parameter drift, the method of linking, and the model used for missing data did not affect findings as much.

To conclude, although our analyses revealed some variability in effect sizes across the specifications, all results were significant and pointed in the same direction.

References

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16. https://doi.org/10.1037/a0022640

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In *New Horizons in Testing* (pp. 31–49). Elsevier. https://doi.org/10.1016/B978-0-12-742780-5.50010-X

Haberman, S. J. (2009). Linking parameter estimates derived from an item response model through separate calibrations. *ETS Research Report Series*, *2009*(2), i-9. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

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

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In Search of Matthew Effects in Reading. *Journal of Learning Disabilities*, 49(5), 499–514. https://doi.org/10.1177/0022219414559974

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional irt models. *Educational and Psychological Measurement*, *66*(1), 63–84. https://doi.org/10.1177/0013164404273942

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die bedeutung der itemauswahl und der modellwahl für die längsschnittliche erfassung von kompetenzen [relevance of item selection and model selection for assessing the development of competencies: The development in reading competence in primary school students]. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, *43*(4), 213–227. https://doi.org/10.1026/0049-8637/a000052

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.2694998

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 11(5), 702–712. https://doi.org/10.1177/1745691616658637

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*(3), 461–475. https://doi.org/10.1007/BF02294761

Taherbhai, H., & Seo, D. (2013). The philosophical aspects of irt equating: modeling drift to evaluate cohort growth in large scale assessments. *Educational Measurement: Issues and Practice, 32*(1), 2–14.

Supplement D: Residualized change scores

Comparisons of change scores from different achievement groups are prone to suffer from regression toward the mean, especially when the first achievement measurement is used for group categorization (Campbell & Kenny, 2003). That is, when measuring a construct at two or more points in time, it is likely that those with more extreme values during the first measurement will show less extreme values at later measurement points. This might lead researchers to conclude that compensation effects are present even though the lower relative growth for high-achieving students is a methodological artefact produced by regression toward the mean. It has been suggested that correcting a change score for regression to the mean can be done by comparing *residualized change scores* (Campbell & Kenny, 2003). Residualized change scores have been recommended as a robustness check for effects of regression toward the mean (Campbell & Kenny, 2003; Heimendinger & Laird, 1983). A residualized change score denotes the change the person would have shown if everyone had started out equal and regression artifacts had not affected the individuals differentially depending on their prior score. Thus, it can be interpreted as the difference between a person's observed and predicted score at Time 2. These residualized change scores are each person's residuum of the change variable and an estimate of how far they deviate from their expected change, given regression toward the mean.

We calculated residualized change scores between the first two measurement points according to Campbell & Kenny (2003) using the standard formula $Y - b_{yx}(X - M_X) - M_Y$. Thus, we calculated the residualized change score for each person on the basis of the total group's observed mean and slope parameters. These residuals were pooled within the high-achieving and the other group and this mean residualized change was then compared across groups to determine the mean change in the groups after controlling for regression artifacts.

If high achievers were to have a residualized change below zero, this would mean that their increase in achievement was smaller than expected. This would suggest that a compensating effect exists over and above the regression toward the mean effect. Therefore, we compared each achievement group's mean value on the resulting residualized change variable. The mean residualized change score for initially high-achieving students in reading was -0.15, whereas the mean residualized change score for other students was 0.01. However, this difference between the two groups' scores was only marginally significant, $t(170.06) = 1.86$, $p = .06$. This indicates that the high achievers' deviation from the expected change was likely smaller than the other group's deviation, and thus, their change in achievement might be attributable to regression toward the mean to some extent, but above this effect, there was still a compensation process.

Table II-D1

Residualized Change Scores for Unconditioned Piecewise Model

	Reading		Mathematics	
	M	(SD)	M	(SD)
Competence at t1 (total sample)	0.00	(0.64)	-0.01	(0.81)
Competence at t2 (total sample)	0.71	(0.86)	0.72	(1.00)
Slope 1 (total sample)	0.70	(0.00)	0.73	(0.00)
Residualized change (high achievers)	-0.15	(0.93)	-0.06	(0.84)
Residualized change (other students)	0.01	(0.74)	0.00	(0.83)

Note. The difference in the residualized change scores in reading was marginally significant (p = .06).

In mathematics, the picture looked different. The residualized mean change score for high achievers was -0.06, whereas the mean change score for the other students was 0.00, and the difference was not significant, $t(189.68) = 0.84$, p = .40. Thus, the residualized change scores signified no differences between achievement groups in growth in mathematics between Grades 5 and 6. The change score analyses suggest that the early differences in growth between the two groups may stem partly from regression toward the mean. However, as Rogosa (1982) pointed out, residual change scores have to be interpreted with great caution. Residualized scores cannot be seen as a corrected measure for change, as the proportion of change discarded by controlling for differences in T1 still contains genuine and important change. Therefore, it can be concluded that in reading, the compensation effect seems, to some extent, to exist over and above the effect of regression to the mean. For mathematics, this conclusion cannot be made from these analyses. Future research might find better ways to deal with the problem of regression to the mean.

References

Campbell, D. T., & Kenny, D. A. (2003). *A primer on regression artifacts* (Reprinted.). *Methodology in the social sciences*. New York: Guilford Press

Heimendinger, J., & Laird, N. (1983). Growth changes: Measuring the effect of an intervention. *Evaluation Review, 7*(1), 80–95. https://doi.org/10.1177/0193841X8300700105

1.3. Zusammenfassung und Zwischenfazit aus Artikel II

Der vorangegangene Beitrag liefert keine Evidenz dafür, dass die Leistungsschere zwischen den anfangs Leistungsstärkeren und ihren Mitschülerinnen und Mitschülern sich im Mittel über die Sekundarstufe I vergrößert. Im Gegenteil, die Befunde sprechen für einen Kompensationseffekt zu Beginn der Sekundarstufe mit anschließend paralleler Entwicklung. Was bedeutet dies für die Schulpraxis? Das Gymnasium ist die Schulform in Deutschland, welche in besonderem Maße die Förderung leistungsstarker Schülerinnen und Schüler gewährleisten soll. Dies begründet zumindest teilweise die Legitimität des gegliederten Schulsystems.

Es wurde jedoch bereits in der Vergangenheit darauf hingewiesen, dass Lehrkräfte teilweise nicht alle Kinder ihrem Fähigkeitsniveau entsprechend fördern. So besagt die Steuerungsgruppen-Theorie, dass Lehrkräfte ihren Unterricht am unteren Fähigkeitsdrittel ausrichten (Lundgren, 1972; zit. nach Treinies & Einsiedler, 1996). Dies könnte unter anderem auch mit Gerechtigkeitsüberzeugungen der Lehrkräfte erklärt werden (Schwippert & Walker, 2003). In diesem Fall würden Lehrkräfte sich zu Beginn der Sekundarstufe stärker darauf konzentrieren, das Leistungsniveau innerhalb ihrer Klassen zu homogenisieren, statt die Leistungsstärksten in besonderem Maße herauszufordern, da dies (unerwünschte) Unterschiede verstärken könnte. Als Optimum der Leistungsentwicklung in Schulklassen wird häufig ein hohes Leistungsniveau bei gleichzeitig geringer bzw. abnehmender Leistungsstreuung genannt (Souvignier & Gold, 2006) – ein Muster, welches in bisherigen Analysen allerdings ein geringer Anteil an Klassen zeigte (Baumert, Roeder, Sang & Schmitz, 1986; Helmke, 1988). Häufiger waren abnehmende Streuung bei unterdurchschnittlichem Leistungszuwachs (in Deutsch, 37 % der Klassen) oder überdurchschnittlicher Leistungszuwachs bei zunehmender Streuung (in Mathematik: 32 % und Englisch: 33 %). Reduktionen in der Leistungsstreuung gingen demnach zulasten der Leistungsstärkeren in der Klasse (Baumert et al., 1986). Allerdings ist einschränkend zu konstatieren, dass die entsprechenden Studien bereits in den 1980er Jahren erfolgten. Mehrebenenanalysen könnten hier einen Hinweis darauf geben, in welchem Maße der Kompensationseffekt zwischen Klassen und in welchem Maße er innerhalb von Klassen stattfindet. Im genannten Fall wäre ein großer Anteil der Kompensation innerhalb von Schulklassen verortet.

Die Ergebnisse dieser Studie können allerdings auch anders interpretiert werden: Schülerinnen und Schüler, die auf das Gymnasium wechseln, sollten in der Regel bereits die Leistungsspitze an ihren Grundschulen darstellen. Da bekannt ist, dass in unterschiedlichen Schulklassen und Schulen Lehrkräfte unterschiedliche Leistungsstandards setzen, spiegelt die Varianz der Leistungen nach dem Übergang ans Gymnasium nicht unbedingt Varianz im Lernpotenzial wider, sondern auch

Varianz im Anregungsgehalt des Grundschulunterrichts. (Im IQB-Bildungstrend 2016 zeigte sich, dass in Deutsch und Mathematik etwa ein Fünftel der Varianz in den Leistungen auf Varianz zwischen verschiedenen Klassen zurückzuführen war (Haag & Kohrt, 2017; Wittig & Weirich, 2017).) Damit wäre es zu einem gewissen Grad zu erwarten, dass auch diejenigen, die - relativ zu anderen - geringere Leistungen zu Beginn der Sekundarstufe vorweisen, am Gymnasium zunächst einen Leistungssprung machen können, wenn dort die Förderung auf einem höheren, adäquateren Niveau stattfindet. Dies würde zu einem Muster wie dem in diesen Analysen beobachteten führen. Diese Annahme könnte empirisch gestützt werden, wenn der initiale Leistungszuwachs am Gymnasium in einem stärkeren Maße vom kognitiven Fähigkeitspotenzial oder der Intelligenz (als Maß für das einer Person innewohnenden Lernpotenzials) abhängig wäre.

Im Rahmen dieser kumulativen Arbeit stellt der Beitrag auch ein Beispiel für die Anwendung der Multiversumsanalyse als Instrument zur Prüfung der Robustheit des Ergebnisses über verschiedene Operationalisierungen hinweg dar. Dabei wurde 1) die Operationalisierung der Gruppe der Leistungsstarken, 2) die Skalierung des Testinstruments und 3) der Umgang mit fehlenden Werten variiert. Insgesamt zeigte sich, dass der Effekt über alle Spezifikationen zwischen $d = 0.54$ und $d = 1.14$ in Mathematik und zwischen $d = 0.76$ und $d = 1.59$ im Lesen variierte. Es zeigte sich, dass der Unterschied in den Wachstumsraten zwar etwas geringer wurde, wenn ein größerer Anteil der Stichprobe als leistungsstark eingruppiert wurde, aber die Unterschiede waren immer noch bedeutsam und keine Effektumkehr fand statt. In beiden Fächern war darüber hinaus ein deutlicher Einfluss der Skalierung des Testinstruments deutlich, das betraf die Frage, welche Stichprobe zur Skalierung herangezogen wurde und ob Items auf Grundlage ihrer Passung im Rasch-Modell ausgeschlossen wurden. Damit werden Ergebnisse der Meta-analyse von Pfost et al. (2014), welche über alle Studien hinweg einen deutlichen Einfluss der Eigenschaften des eingesetzten Testinstruments fanden, auch innerhalb des vorliegenden Beitrags durch die Anwendung der Multiversumsanalyse bestätigt.

Literaturverzeichnis

Alexander, K. L., Entwisle, D. R. & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, *23*(2), 171–191. https://doi.org/10.3102/01623737023002171

Alexander, K. L., Entwisle, D. R. & Olson, L. S. (2007). Lasting Consequences of the Summer Learning Gap. *Am Sociol Rev*, *72*(2), 167–180. https://doi.org/10.1177/000312240707200202

Artelt, C., Blossfeld, H.-P., Faust, G., Roßbach, H.-G. & Weinert, S. (2013). *Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (BiKS-8- 14).* Datset, https://doi.org/10.5159/IQB_BIKS_8_14_v2. https://doi.org/10.5159/IQB_BIKS_8_14_v2

Bast, J. & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a dutch longitudinal study. *Developmental Psychology*, *34*(6), 1373–1399.

Baumert, J., Becker, M., Neumann, M. & Nikolova, R. (2009). Frühübergang in ein grundständiges Gymnasium – Übergang in ein privilegiertes Entwicklungsmilieu? *Zeitschrift für Erziehungswissenschaft* [Early transition into the academic track of secondary schooling – Transfer into a privileged learning environment?], *12*(2), 189–215. https://doi.org/10.1007/s11618-009-0072-4

Baumert, J., Roeder, P. M., Sang, F. & Schmitz, B. (1986). Leistungsentwicklung und Ausgleich von Leistungsunterschieden in Gymnasialklassen. *Zeitschrift für Pädagogik*, *32*(5), 639–660.

Ceci, S. J. & Papierno, P. B. (2005). The rhetoric and reality of gap closing: when the "have-nots" gain but the "haves" gain even more. *The American Psychologist*, *60*(2), 149–160. https://doi.org/10.1037/0003-066X.60.2.149

Crawford, C., Macmillan, L. & Vignoles, A. (2016). When and why do initially high-achieving poor children fall behind? *Oxford Review of Education*, *43*(1), 88–108. https://doi.org/10.1080/03054985.2016.1240672

DiPrete, T. A. & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, *32*(1), 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Haag, N. & Kohrt, P. (2017). Mittelwerte und Streuungen der im Fach Mathematik erreichten Kompetenzen. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 169–186). Münster: Waxmann.

Helmke, A. (1988). Leistungssteigerung und Ausgleich von Leistungsunterschieden in Schulklassen: Unvereinbare Ziele? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *20*(1), 45–76.

Kocaj, A., Kuhl, P., Kroth, A. J., Pant, H. A. & Stanat, P. (2014). Wo lernen Kinder mit sonderpädagogischem Förderbedarf besser? Ein Vergleich schulischer Kompetenzen zwischen Regel- und Förderschulen in der Primarstufe. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *66*(2), 165–191. https://doi.org/10.1007/s11577-014-0253-x

(1912). *Lutherbibel*. Zugriff am 14.12.2021. Verfügbar unter: https://bibeltext.com/l12/

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Morgan, P. L., Farkas, G. & Hibel, J. (2008). Matthew Effects for Whom? *Learning disability quarterly : journal of the Division for Children with Learning Disabilities*, *31*(4), 187–198.

Pfost, M., Hattie, J., Dörfler, T. & Artelt, C. (2014). Individual differences in reading development. A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research*, *84*(2), 203–244. https://doi.org/10.3102/0034654313509492

Pfost, M., Karing, C., Lorenz, C. & Artelt, C. (2010). Schereneffekte im ein- und mehrgliedrigen Schulsystem. *Zeitschrift für Pädagogische Psychologie* [Fan spread effects in a tracked and a nontracked school system], *24*(3-4), 259–272. https://doi.org/10.1024/1010-0652/a000025

Prenzel, M., Reiss, K. & Hasselhorn, M. (2010). Förderung der Kompetenzen von Kindern und Jugendlichen. In J. Milberg (Hrsg.), *Förderung des Nachwuchses in Technik und Naturwissenschaft. Beiträge zu den Zentralen Handlungsfeldern* (acatech DISKUTIERT, S. 15–60). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schwippert, K. & Walker, M. (2003). Homogenous and high performing classes: The case of optimal classes. *Studies in Educational Evaluation*, *29*(2), 109–128. https://doi.org/10.1016/S0191-491X(03)00018-X

Souvignier, E. & Gold, A. (2006). Wirksamkeit von Lehrmethoden. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik. Mit 18 Tabellen* (S. 146–166). Berlin, Heidelberg: Springer Medizin Verlag Heidelberg. https://doi.org/10.1007/3-540-33020-8_10

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *XXI*(4), 360–406.

Treinies, G. & Einsiedler, W. (1996). Zur Vereinbarkeit von Steigerung des Lernleistungsniveaus und Verringerung von Leistungsunterschieden in Grundschulklassen. *Unterrichtswissenschaft*, *24*, 290–311. Verfügbar unter: urn:nbn:de:0111-opus-79408

Walberg, H. J. & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*(3), 359–373.

Wittig, J. & Weirich, S. (2017). Mittelwerte und Streuungen der im Fach Deutsch erreichten Kompetenzen. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 153–167). Münster: Waxmann.

Zuckerman, H. (2010). Dynamik und Verbreitung des Matthäus-Effekts. *Berliner Journal für Soziologie*, *20*(3), 309–340. https://doi.org/10.1007/s11609-010-0133-9

4

Die soziale Integration leistungsstarker Schülerinnen und Schüler

4. Die soziale Integration leistungsstarker Schülerinnen und Schüler

Die sozialen Erfahrungen, die Kinder und Jugendliche in der Schule machen, sind von großer Bedeutung sowohl für ihr emotionales Wohlbefinden als auch für ihre persönliche und schulische Entwicklung (Ladd, 2007; Osterman, 2000). Aus diesem Grund wurden in der Vergangenheit bereits sowohl Prädiktoren (z.B. Alter, Geschlecht, Geschlechterrollenkonformität, Schulleistungen, Persönlichkeitseigenschaften, soziale Kompetenzen, Eigenschaften der Interaktionspartner oder Peergruppencharakteristika; Brown, 2004) als auch Folgen unterschiedlicher sozialer Erfahrungen von Schülerinnen und Schülern untersucht (Gifford-Smith & Brownell, 2003). Im vorliegenden Beitrag befassen wir uns spezifisch mit der sozialen Einbettung leistungsstarker Schülerinnen und Schüler in ihren Schulklassen.

4.1. Theoretischer Hintergrund

Die Entwicklung einer persönlichen und sozialen Identität wird häufig als eine vornehmliche Entwicklungsaufgabe des Jugendalters genannt (Verhoeven, Poorthuis & Volman, 2019). In dieser Zeit suchen Jugendliche Anschluss an eine soziale Gruppe, deren Mitglieder ein gemeinsames Verständnis davon entwickeln, was es heißt, dieser Gruppe anzugehören und was diese Gruppe positiv von anderen abhebt. Diese Gruppennormen werden dann situativ aktiviert, um eine das eigene Verhalten zu steuern und das Verhalten anderer zu bewerten (Turner & Reynolds, 2012). Der Wunsch nach positiver sozialer Distinktheit kann dazu führen, dass Gruppenmitglieder, welche in bedeutsamen Eigenschaften von der Gruppe abweichen, dazu motiviert werden, sich der Gruppe anzunähern oder sie werden von der Gruppe ausgeschlossen. Konformität hingegen wird mit sozialem Status belohnt (Festinger, 1954). Die Konformität mit Peernormen hat daher in diesem Alter eine besondere Relevanz für die sozialen Interaktionen zwischen Peers.

Da im Unterricht täglich Leistungsvergleiche angestellt und Leistungsverhalten durch die Lehrkräfte sanktioniert oder bestärkt wird, erhalten diese eine besondere Salienz im Schulalltag. Die Art und Weise, wie die Leistungsnormen Einfluss auf die sozialen Konstellationen in der Schulklasse nehmen, wurde in unterschiedlichen Forschungsbereichen untersucht. So beschäftigt sich ein großer Forschungsbereich mit der Rolle von Peerkulturen, die die von den Erwachsenen vorgegebenen Werte in Frage stellen (Pugh & Hart, 1999). Dabei wird die Annahme getroffen, dass die Ablehnung bisher gültiger Normen und die Entwicklung einer jugendlichen Gegenkultur Strategien zur Ausbildung einer eigenen Identität darstellen. Schüler und Schülerinnen, die von dieser oppositionellen Peernorm abweichen und stattdessen leistungsorientiert sind und damit die Ansprüche von Eltern und Lehrkräften erfüllen, werden als „Streber“ verunglimpft (Rentzsch, Schröder–Abé & Schütz, 2013). Unterstützung erfährt diese Deutung durch Untersuchungen, die zeigen, dass unter Jugendlichen sozialer Status durch deviantes Verhalten den Lehrkräften

gegenüber gewonnen werden kann (Brown & Larson, 2009). Andererseits zeigen größere Untersuchungen, dass Klassen, in denen tatsächlich leistungsfeindliche Peernormen vorherrschen, in der Minderzahl sind (Kruse & Kroneberg, 2020).

Eine differenziertere Sicht auf den Zusammenhang zwischen Leistungsverhalten und dem Peerkontext entsteht in Forschungsarbeiten, die sich mit den Beziehungen zwischen Geschlechterrollen und Leistungserwartungen beschäftigen. Hier ist die Annahme, dass gesellschaftliche Rollenerwartungen an Männer bzw. Frauen existieren, mit denen sich Jugendliche auseinandersetzen und denen sie in der Regel versuchen, zu entsprechen (Kessels, Heyder, Latsch & Hannover, 2014). Das Erbringen von Leistungen in unterschiedlichen Schulfächern wird im Einklang oder Dissens mit diesen Geschlechterstereotypen wahrgenommen (Kessels, 2005). Hohe Leistungen werden insbesondere dann als kritikwürdig wahrgenommen, wenn sie mit dem Geschlechterrollenbild nicht konform sind (Kessels et al., 2014).

Gemeinsam ist einem großen Teil dieser Forschung, dass überwiegend injunktive Normen untersucht (z.B. Kessels, 2005; Workman & Heyder, 2020) oder Vignetten eingesetzt werden, um Einstellungen gegenüber hypothetischen Schülerinnen und Schülern zu bestimmen (z.B. Händel, Vialle & Ziegler, 2013; Hannover & Kessels, 2004; Kessels, 2005). Tatsächliche soziale Beziehungen unter Schülerinnen und Schülern sind meist nicht teil der Untersuchung. Damit eignen sich diese Studien nur bedingt, um die tatsächliche Einbindung Leistungsstarker in ihren sozialen Klassenkontext zu beurteilen, auch wenn sie hilfreich in der Ableitung von Hypothesen über den Zusammenhang zwischen Leistungsstärke und sozialer Integration sind. Dort, wo konkretes Verhalten gegenüber Mitschülerinnen und Mitschülern untersucht wird, beschränkt sich dieses meist auf Viktimisierung (z.B. Bergold, Kasper, Wendt & Steinmayr, 2020; Boehnke, 2008; Gest, Graham-Bermann & Hartup, 2001; Pelkner & Boehnke, 2003).

Dieser Forschung zu gruppenbasierten Einstellungen und Verhaltensweisen, welche eher in sozialpsychologischer Tradition stehen, steht ein Forschungsfeld gegenüber, in welchem vielfältige Formen schulischer Peerbeziehungen untersucht werden (Zander, Kreutzmann & Hannover, 2017). Peerbeziehungen können dabei viele unterschiedliche Aspekte umfassen, als Beispiele seien hier Freundschaft, Akzeptanz, das Eingebettetsein in Netzwerke, oder sozialer Status genannt (Abrams & Killen, 2014; Cillessen, Schwartz & Mayeux, 2011; Gifford-Smith & Brownell, 2003; Ladd, Kochenderfer & Coleman, 1997; Youniss & Haynie, 1992). Die soziale Integration ganz allgemein betrifft die Einbettung einer Person in ihr soziales Umfeld. Eine Person kann in verschiedenen Systemen (z.B. Klassen-/ Schulkontext; Freizeitkontext) und Systemebenen (Dyade, Peergruppe, Klassensystem) und bezüglich verschiedener Facetten - die

stärker behavioral oder stärker affektiv gefärbt sein können (z.B. Freundschaft, Akzeptanz, Teilnahme an gemeinsamen Aktivitäten, Zufriedenheit mit sozialen Beziehungen) - unterschiedlich integriert sein. Die Messung und Quantifizierung sozialer Beziehungen kann, je nach untersuchter Facette, Zielstellung und Methodik der Untersuchung sehr unterschiedlich erfolgen. Ein wichtiges Paradigma bei der Untersuchung sozialer Beziehungen stellen soziale Netzwerkanalysen dar, welche besonders den quantitativen Aspekt sozialer Beziehungen in den Mittelpunkt stellen. Auf Basis von Nominierungsnetzwerken in sozialen Gruppen können dabei eine Vielzahl an Indikatoren berechnet werden. Eine stärker qualitative Beschreibung sozialer Integration hingegen kann beispielsweise mit psychometrischen Skalen zur selbst eingeschätzten Integration oder zur Zufriedenheit mit Beziehungen erfolgen. Bisherige Untersuchungen zur sozialen Integration leistungsstarker Schülerinnen und Schüler ergaben, dass bessere Leistungen in der Regel auch mit größerer sozialer Akzeptanz einhergehen (Wentzel, Jablansky & Scalise, 2021).

Auch wenn verschiedentlich darauf hingewiesen wird, dass verschiedene Messinstrumente und Indikatoren unterschiedliche Aspekte sozialer Beziehungen betreffen, fehlt eine übergreifende Konzeption sozialer Integration, welche die unterschiedlichen Aspekte einordnet und miteinander in Beziehung setzt. Koster, Nakken, Pijl und van Houten (2009) unternahmen eine Literaturstudie, bei der sie die Konzepte sozialer Integration in Studien aus dem Bereich der Sonderpädagogik herausarbeiteten. Dabei identifizierten sie vier distinkte Bedeutungsschwerpunkte: Beziehungen, Interaktionen, Akzeptanz und subjektive Integration. Für jede dieser Facetten sozialer Integration nannten sie Beispiele (s. Abbildung 5). Es stellt sich die Frage, ob diese Einteilung ein brauchbares Instrument darstellt, um die soziale Integration leistungsstarker Schülerinnen und Schüler systematisch zu untersuchen.

4.2. Ableitung der Forschungsfragen und methodisches Vorgehen

Soziale Integration als Konzept ist vielgestaltig. Sie wird sowohl subjektiv als auch objektiv, sowohl in dyadischen Beziehungen als auch in Gruppenkontexten begriffen und gewinnt seine Bedeutung aus affektiven Folgen und instrumentellen Nutzen (Zander et al., 2017). Bossaert, Colpin, Pijl und Petry (2013) und Koster et al. (2009) unterscheiden Beziehungen, Interaktionen, Akzeptanz und subjektive Integration als vier Facetten der sozialen Integration. Sie beanspruchen dabei nicht, die soziale Integration als Konzept vollständig abgebildet zu haben. Stattdessen sehen sie die Analyse als Ausgangspunkt weiterer Forschung, welche die Beziehung der Facetten untereinander sowie ihre differenziellen Zusammenhänge mit Auswirkungen und Einflussfaktoren untersucht. Auch wenn sie in ihrer Analyse Beispiele für die einzelnen Facetten nennen, steht doch bei der Vielfalt der in der bisherigen Forschung genutzten Maße für soziale Integration die Frage

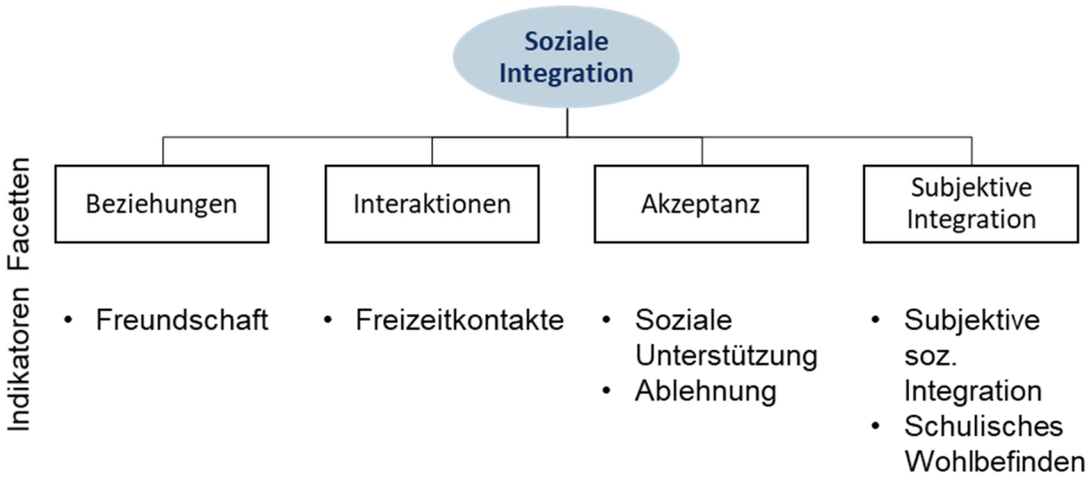


Abbildung 4.1: Facetten der sozialen Integration nach Bossaert et al. (2013) und Koster et al. (2009).

der Operationalisierung bzw. Zuordnung von Messinstrumenten und dahinterstehenden Konstrukten im Raum. Bezugnehmend zu weiterer Forschung, die sich mit sozialen Beziehungen beschäftigt, ordnen wir als Erstes unterschiedliche Indikatoren sozialer Integration diesen Facetten zu.

In einem zweiten Schritt untersuchen wir den Zusammenhang von Leistungsstärke und diesen verschiedenen Facetten sozialer Integration. Dabei stellen wir unterschiedliche Indikatoren für Leistungsstärke nebeneinander. Über diese verschiedenen Operationalisierungen von Leistungsstärke und sozialer Integration hinweg untersuchen wir den Zusammenhang beider Bereiche. Dieser Schritt ist noch nicht konfirmatorisch, sondern explorativ und daher im Entdeckungszusammenhang zu verorten, kann also weiterer Theorieentwicklung als Basis dienen. Es stellt sich die Frage, ob der Zusammenhang zwischen Leistung und sozialer Integration sich zwischen den Facetten und genutzten Indikatoren unterscheidet. Dies würde dafür sprechen, die Struktur des Konstrukts soziale Integration künftig genauer zu untersuchen.

Im dritten Schritt wenden wir uns Fragestellungen zu, die wir aus bisherigen Theorien ableiten. Insbesondere soll erstens geprüft werden, ob möglicherweise ein nichtlinearer Zusammenhang die Unterschiede zwischen sozialen Netzwerkstudien - die einen positiven Zusammenhang zwischen Leistung und sozialer Integration finden - und Stereotypenforschung - die häufig eine eher negative Sicht von leistungsstarken Schülerinnen und Schülern feststellt – erklären kann. So stellen die meist eingesetzten Vignetten „hypothetischer“ neuer Klassenkameradinnen und Klassenkameraden Extrembilder dar, während der positive Zusammenhang zwischen Leistung und Integration meist als linearer Zusammenhang auf Grundlage des gesamten Leistungsspektrums berechnet wird. Auch Studien, die sich mit Viktimisierung Leistungsstarker

beschäftigen, legen durch die Betrachtung von Extremgruppen keinen linearen Zusammenhang zwischen Leistung und sozialen Erfahrungen an (z.B. Bergold, Kasper, Wendt & Steinmayr, 2020).

Zweitens wird die Rolle des Geschlechtes im Zusammenhang zwischen Leistungsstärke und sozialer Integration untersucht. Bisherige Forschung kommt zu dem Schluss, dass von Schülerinnen und Schülern wahrgenommene Geschlechterrollen für die schulische Leistung und im Peerkontext eine große Rolle spielen (Kessels, Heyder, Latsch & Hannover, 2014). Insbesondere zeigt sich in Untersuchungen zu Stereotypen, dass Einstellungen gegenüber (hypothetischen) Mitschülerinnen und Mitschülern, deren Leistung nicht dem wahrgenommenen Geschlechterrollenstereotyp entspricht, negativer sind (Kessels, 2005). Aus diesen Befunden leitet sich die Frage ab, ob dieses Ergebnis auch repliziert werden kann, wenn nicht Einstellungen gegenüber Vignetten, wie sie in der Forschung häufig verwendet werden, sondern die Einstellungen gegenüber konkreten Klassenkameradinnen und Klassenkameraden mit unterschiedlichen Leistungsprofilen betrachtet werden.

Um diese Fragestellungen zu bearbeiten, werden im Beitrag unterschiedliche Ansätze kombiniert. Zunächst wird eine Multiversumsanalyse durchgeführt, um Zusammenhänge zwischen verschiedenen Facetten von sozialer Integration und verschiedenen Operationalisierungen und Teilbereichen von Leistung zu explorieren. Dabei wird für jede Kombination eine Korrelation bestimmt, welche dann in aufsteigender Größe - gemeinsam mit ihren zugehörigen Spezifikationen - auf einer Kurve angeordnet werden, um Muster identifizieren zu können.

Die weiteren Fragestellungen beziehen sich nicht mehr auf das gesamte Multiversum an Operationalisierungen von Leistung. Stattdessen wird der Notenschnitt als Maß für Leistung herangezogen. Der Grund hierfür ist, dass Noten ein sichtbares Abbild von Schulleistungen darstellen, welche von Klassenkameraden und -kameradinnen vermutlich am ehesten als Richtschnur genutzt werden, um soziale Vergleiche innerhalb der Klasse durchzuführen. Außerdem lagen Noten in unterschiedlichen Leistungsdomänen (sprachlich und mathematisch-naturwissenschaftlich) vor.

Um die Hypothese zu untersuchen, ob ein nichtlinearer Zusammenhang zwischen Leistung und sozialer Integration besteht, wurde zunächst ein two-lines test von Simonsohn (2018) durchgeführt, der das Vorhandensein eines Cut-wertes testet, an welchem ein Scheitelpunkt existiert, der einen Zusammenhang in zwei gegengesetzte lineare Abschnitte trennt. Als zweites wurden Regressionen von den Facetten sozialer Integration auf Leistung durchgeführt und beurteilt, ob der quadratische Leistungsterm signifikant ist. Für den Zusammenhang zwischen

Leistung und Geschlecht wurde in diese Regression ein Interaktionsterm eingeführt, welcher die Interaktion zwischen mittlerer Leistung und Geschlecht beinhaltete. Um schließlich die Hypothese zur Fachbezogenheit von sozialer Integration, Geschlecht und Leistung zu untersuchen, wurde eine Variable berechnet, welche angab, ob eine Person ein Leistungsprofil zeigte, welches einem bestimmten Geschlechterstereotyp entsprach. Die Studie wurde präregistriert, die Präregistrierung sowie der Analysecode sind verfügbar unter https://osf.io/hu2n8/ .

III

The Social Integration of High-Achieving Secondary School Students in Their Classroom: No Evidence for an Interaction with Gender and School Subject

Claudia Neuendorf

Malte Jansen

Dies ist die eingereichte Fassung eines Manuskripts, das inzwischen beim Journal of Educational Psychology zur Publikation angenommen wurde. Die finale Version des Artikels wird nach ihrer Publikation unter der folgenden DOI verfügbar sein: https://doi.org/10.1037/edu0000778

Abstract

Prior research has found that student achievement is positively related to students' social position in class. However, negative stereotypes about high academic achievers prevail among secondary school students, suggesting that higher achievers might be less well integrated socially. These stereotypes especially target academically high-achieving boys, students achieving highly in mathematics and sciences and students performing well in opposite-gender stereotyped subjects. This article tries to link these stereotypes to actual social integration measured by self-report and social network nominations. It tests whether there is an inversely u-shaped relationship between achievement and social integration, and whether high performance in mathematics, physics or in opposite gender stereotyped ways might be associated with lower levels of social integration. Using data from a German large-scale assessment study with about 45,000 ninth-grade students, we conducted a multiverse analysis on correlations between multiple measures of achievement (grades, test scores and achievement self-concept) and multiple facets of social integration (friendship, acceptance, contact, and subjective integration). Further, we evaluated the shape of the relationship using the two-lines test. Only the facet of friendship followed the hypothesized shape. Finally, in a series of regression analyses, we found no support for an interaction between general achievement level and gender and no interactions between subject of achievement and gender, except for the facet of acceptance (being asked for help by peers). We conclude that high-achieving students are not at a higher risk for social exclusion and stereotypes seem not to align with actual social relationships present in secondary school classes in Germany.

Keywords: High Achievement; Stereotypes; Peer Relationship; Social Networks; Secondary School

Educational impact and implications statement

Our study was in search of patterns indicating stereotyped choice of social interaction partners in secondary school classrooms. Specifically, we were interested in the social integration of high achievers. Overall, higher achievers are better integrated in their classroom. Moreover, we did not find any signs that boys and girls performing well in opposite gender stereotyped subjects (Language and Biology vs. Mathematics and Physics) were liked less by their peers.

Introduction

Students' social experiences in school are highly important for their emotional well-being as well as their school-related behavior (Osterman, 2000). For example, Students who are well liked by their peers show positive adjustment and prosocial behavior (Brown & Larson, 2009), while children without friends have been shown to report lower social, emotional and academic adjustment (Wentzel et al., 2004). Thus, the importance of social relationships at school has been widely acknowledged and predictors at the individual and the group level on the one hand as well as outcomes of social experiences on the other hand have been studied (Gifford-Smith & Brownell, 2003). One factor which is repeatedly found to correlate positively with popularity and social acceptance is academic achievement. However, this mostly consistent finding that students who perform better at school are also more accepted (Wentzel et al., 2021) is at odds with the pervasive stereotype of the highly achieving but socially excluded student (as is implicated by the well-known terms nerd, geek, teacher's pet or, in the German case, *Streber*; Boehnke, 2008; Rentzsch et al., 2013). Our study aims to disentangle these seemingly contradictory ideas by studying the relation between achievement and social integration from multiple perspectives. We assume that the phenomenon of social exclusion of high achievers—if it indeed exists—can be best detected and investigated by 1) using non-linear transformations of achievement (e.g., is there a tipping point at which higher achievement does not lead to higher social inclusion, but becomes detrimental?), 2) considering various indicators of social acceptance, and 3) considering various indicators of achievement. Following up on earlier studies showing gender differences in the experience of social relationships between highly achieving boys and girls (Bergold et al., 2020) and gender stereotyped domain-specific evaluations of high achievers (Kessels, 2005), we further explore the link between gender, subject of achievement and social integration of high achievers in their classes.

Social Integration

According to Koster et al. (2009) and Bossaert et al. (2013), who have conducted reviews on the use of the terms social integration, social acceptance and social inclusion in studies on children with special educational needs, social integration of students can be measured on four (interrelated) dimensions: 1) having mutual friendships or positive relationships with others, 2) having contacts or positive interactions with others, 3) being accepted by peers and 4) perceiving oneself as socially accepted by peers. These different dimensions relate to different levels on which relationships are studied. Self-perceived social integration can be measured on an individual level, whereas friendships and interactions take place at a dyadic level. Finally, peer group social acceptance is measured by aggregating individual ratings at the group level. Research has suggested that these

dimensions are, at least to some extent, separate (Youniss & Haynie, 1992): friendships, self-perceived social integration and social acceptance in the peer group each uniquely predict psychosocial outcomes (Gifford-Smith & Brownell, 2003; Krawinkel et al., 2017; Ladd et al., 1997). Moreover, as perceptions between individuals regarding their social relationships differ (Brown & Larson, 2009; Gifford-Smith & Brownell, 2003), self-perceived social acceptance does not have to mirror actual peer acceptance (Putarek & Keresteš, 2016).

Because of these considerations, researchers studying social integration should be clear on which aspect of integration they are targeting at a conceptual level (Bossaert et al, 2013; Koster et al., 2009). Especially, peer group acceptance can take very different forms. Sometimes, social standing in the peer group is conceptualized as the desirability as an interaction partner or friend (*sociometric popularity*), sometimes, it is framed as having a high peer-perceived social status. These two forms of popularity are very different in their implications and students with a high social status are not always students with many friends or students who are liked by others (Vörös et al., 2019), because status is sometimes gained by anti-social behavior (Brown & Larson, 2009; Gifford-Smith & Brownell, 2003). In this article, we refer to social integration as an umbrella term to cover all of the sub-dimensions, that is, friendship, contacts, acceptance, and self-perceived acceptance.

Different approaches to the measurement of social integration have been developed over time which lend themselves to the operationalization of the dimensions of social integration. Broadly, sociometric methods, which rely on students submitting information about their classmates, and psychometric methods, which rely on self-reports of social integration, can be distinguished. Mutual friendships and peer acceptance are typically measured by sociometric methods (nomination or rating scales), while self-perceived social integration is assessed with self-report measures. In the current study, we use sociometric as well as psychometric methods to examine different dimensions of social integration and how they relate to academic achievement.

Gender and Social Integration

When studying social integration at school, it is important to acknowledge the gendered social experiences of students. Specifically, same-sex relationships are much more pervasive than cross-gender relationships (Feiring & Lewis, 1991; Maccoby, 1990). Gender differences in social behavior and experiences of adolescents might influence the sub-dimensions of social integration of girls and boys differentially.

Regarding friendships, research has found that boys typically have a larger social network, while girls have more intimate relationships to their peers (Cairns et al., 1998; Gifford-Smith &

Brownell, 2003; Rose & Rudolph, 2006; Rueger et al., 2008; Youniss & Haynie, 1992). Youniss and Haynie (1992) point out that in general, most boys and girls have at least one significant same-sex friend and the sheer network size might not be a predictor of healthy development (see also Benenson, 1990). This matches with the finding that despite these differences in network size, generally, there are no differences in friendship satisfaction between girls and boys (Rose & Rudolph, 2006).

Regarding interactions with peers, these seem to occur with similar frequency within boys' peer groups and within girls' peer groups, although girls' interactions tend to be longer (Rose & Rudolph, 2006). However, there are differences between boys and girls regarding the prevalence of victimization: boys are more often subjected to bullying (Olweus, 1991).

Considering acceptance, there exists only limited research on gender differences (Rubin et al., 2006). Some early works have shown that different child characteristics are important in evaluating peers, but few of them predict popularity differently for boys and girls (Rose et al., 2011). For example, in a study by Benenson (1990) peer acceptance (as measured by peer ratings of desirability as a play partner) was related to the size of a clique a boy belonged to. For girls, this relation between acceptance and individual network size was weaker. Finally, perceived acceptance has been reported to be higher for girls than for boys (Rose & Rudolph, 2006; Rueger et al., 2008).

Achievement and Social Integration

The interplay of achievement and social integration has been a topic of different strands of research and has been studied from different viewpoints. Firstly, there are studies which are concerned with achievement as either an antecedent and/or a consequence of peer acceptance. This research has recently been compiled in a meta-analysis by Wentzel et al. (2021), overall concluding a medium-sized positive link between achievement and acceptance. This pattern has been explained from different theoretical perspectives. For example, students high in achievement typically also show higher levels of prosocial behavior (Becherer et al., 2017) and higher levels of agreeableness (Freund-Braier, 2009; Köller & Baumert, 2017). This finding is especially pronounced for students high in language arts subjects (Hannover & Kessels, 2004). Becherer et al. (2017) suggest that prosocial behavior is predictive of positive relationships between students and this social acceptance might facilitate academic achievement. Although the exact mechanisms linking social acceptance and achievement are not yet well-known, it is discussed whether prosocial behavior is an outcome which is targeted by teachers' instruction and thereby contributes to good grades at the same time as it is conducive of positive relationships. Another argumentation is that learning

is more effective when students can give each other academic support. Research has further shown that study relationships might evolve out of friendships (Stadtfeld et al., 2019). These explanations target individual relationships between students as a measure of acceptance. Other theories focus on social status as a measure of social integration, instead. Thus, Social Referencing Theory (Feinman, 1992) proposes that children use adults as a social reference. This means that teacher reactions towards different types of student behavior are registered by students who then learn which behavior is appropriate (for example, showing high achievement). Students who are openly praised by their teachers might rise in status, thereby becoming a more desired interaction partner (Huber, 2011). However, as students grow older, adult norms become less important to them and status may be gained by other types of behavior (De Laet et al., 2014).

This observation leads to the second line of research: It sheds light on the conditions under which achievement is related to peer acceptance. Here, classroom norms play a dominant role. Within this line of argument, behavior which is normative in a given context is rewarded with status attribution or social preference by peers. The normativity of a certain behavior can be conceptualized in terms of descriptive norms, for example the classroom level of achievement (as in Bond et al., 2017; Boor-Klip et al., 2017). Consistent with the individual-group-similarity hypothesis (Gifford-Smith & Brownell, 2003), higher achievers have been found to be more popular in classrooms with high achievement standards or in academically bound school tracks as opposed to vocational tracks (Bond et al., 2017; Boor-Klip et al., 2017; Kruse & Kroneberg, 2020; Meijs et al., 2010; Palacios et al., 2019; van Houtte, 2006). On the other hand, researchers study peer influence by injunctive norms, that is, which behavior is perceived to be approved or disapproved of by peers. This research touches on issues like peer pressure, stereotypes and conformity expectations experienced by students. It is this kind of research which finds evidence for negative peer effects on high achievers (Händel et al., 2018).

These different findings—a positive relationship between achievement and social acceptance on the one hand and attenuated social integration of high achievers which are perceived as “different” on the other hand—might be further differentiated by exploring the way achievement and social integration are operationalized within both approaches.

Regarding achievement, a dominant measure seems to be classroom grades or GPA. However, standardized tests have also been used by a couple of studies (Bergold et al., 2020; Wentzel et al., 2021; Wolter & Seidel, 2017). In comparison, effects have been larger in size when grades are used compared to standardized test scores (Wentzel et al., 2021; Wolter & Seidel, 2017). Neuendorf et al. (submitted) present a diversity of ways to define high academic achievement and

argue for directly comparing the variability in effects which is due to different operationalizations of this construct. Specifically, different measures (like grades, tests or self-assessments), different cut-offs for defining which achievement is considered as "high", different reference norms (classroom, sample or population norm) and different domains of achievement (e.g., math vs languages) can be chosen. In this vein, Wolter and Seidel (2017) have studied the relationship between self-perceived popularity and different indicators of achievement. They found the grade in mathematics but not in German to be positively related to social integration, while the students' academic self-concept (which could be argued to reflect self-rated academic ability) in German but not in mathematics was predictive of social integration. Achievement as assessed with standardized tests was not related to social integration—neither in mathematics, nor in German. Thus, the measure used for achievement makes a difference in determining the role of achievement in the social experiences of students. Neuendorf et al. (submitted) further suggest that in studies focusing on the social dimensions in the lives of high achievers, achievement measures which capture the social dimension might be most appropriate, by applying a classroom level comparison standard and possibly using measures which are visible to peers, like grades.

Regarding social integration, it has also been argued that a comparability of results is difficult in the face of the different methods that exist to measure social networks and social relationships (Cairns et al., 1998). Again, giving a theoretical rationale for choosing a certain operationalization and clarifying the meaning of these indicators is important. A helpful basis for thinking about different measures of social integration is given in the Bossaert et al. (2013) outline of different key themes implicated in the construct social integration. They suggest studying whether the different key themes can be empirically validated and differential effects on outcome measures can be identified. Correlations between different indicators of social integration have been medium-sized at most (Ladd et al., 1997; Meijs et al., 2010), moreover, Gifford-Smith and Brownell (2003) as well as Vörös et al. (2019) have argued that the overlap in measurement methods for different aspects of social integration might exaggerate the similarity of effects between these concepts. For example, some researchers use reciprocal friendship nominations to measure the number of friends and at the same time use incoming friendship nominations as a measure of sociometric status or acceptance. Studies included in the Wentzel et al. (2021) meta-analysis had measured acceptance by different sociometric approaches, showing the largest effects when a nomination procedure was used and liked most and liked least nominations for every student were subtracted from each other, while other methods, like having students rate every classmate on a social acceptance scale yielded slightly lower effect sizes. In a study by Ladd et al. (1997), the relationship between academic achievement and peer acceptance was strongest

(r = .48), followed by number of friends (r = .37) and loneliness (r = -.34). By contrast, having a very best friend, social dissatisfaction and victimization were less strongly associated with achievement.

To summarize, there seems to be an overall positive relationship between achievement and social integration. The strength of this relationship possibly varies by the sub-dimension of social integration that is in focus and by the measure of achievement used.

Achievement, Gender and Social Integration

In discussions on the relationship between high achievement and social integration, a topic that frequently pops up is the role of gender. In their *Interests as Identity Regulation Model (IIRM),* Kessels et al. (2014) have proposed that students display certain behavior at school as a means of identity regulation. During adolescence, gender becomes an important aspect of students' social selves (Hannover & Zander, 2020; Maccoby, 1990; Mayeux & Kleiser, 2020). Gender stereotypes are used by students (and teachers: Jones & Myhill, 2004) as an orientation in classifying behavior, for example academic engagement in certain subjects, with regard to gender prototypicality. Cultural notions of masculine characteristics include rationality, logical thinking and interest in technology, while femininity is associated with socially oriented behavior, which requires communication skills and with preference for reading (Visser, 1996). Thus, engaging in mathematics and sciences is largely associated with masculinity (Archer et al., 2013; Kessels, 2005), while achievement in language and arts subjects is perceived as more appropriate for girls (Kessels, 2005; Lummis & Stevenson, 1990; Muntoni & Retelsdorf, 2019). Having thus acquired an image of a school subject, students may choose behavioral options in accordance with their social identity. Studies on achievement gaps between boys and girls focus on this mechanism as an explanation of domain-specific achievement differences between boys and girls. Specifically, on average, girls fare better in the language domain whereas some, but not all studies find (small) advantages for boys in the STEM domains (Lindberg et al., 2010; OECD, 2019). However, it should be noted that differences in domain-specific motivational characteristics such as academic self-concepts, interests and achievement emotions are much stronger and more stereotypical than differences in achievement (Frenzel et al., 2007; Gaspard et al., 2015; Jansen et al., 2014; Wang & Degol, 2013). Aside from differences in preferred subject domain, research has also suggested that overall, striving for academic success in general is perceived as more feminine (Jones & Myhill, 2004; Kessels et al., 2014).

An aspect which may exacerbate gendered academic behavior (that is, engagement in school or in certain subjects) is peer pressure experienced by students, which may differ between different

cultures (Boehnke, 2008) and between school contexts (Legewie & DiPrete, 2012). Peer pressure operates on the fear of being socially excluded from a peer group. The premise seems to be that students who do not conform to gender norms will be despised by their peers (Festinger, 1952; Jetten & Hornsey, 2014; Kessels et al., 2014; Masters et al., 2021).

However, while these hypotheses—which can be derived from the Subjective Group Dynamics Theory (Abrams et al., 2003)—have received some support by empirical studies (Bergold et al., 2020; Kessels, 2005; Kleiser & Mayeux, 2021; Smith & Leaper, 2006; Workman & Heyder, 2020), other studies do not confirm gender differences in the relation of (domain-specific) achievement and social integration (Händel et al., 2013; Händel et al., 2018; Rose et al., 2011; Vannatta et al., 2009; Wentzel et al., 2021; Wolter & Seidel, 2017).

The present study

The present study examines the social integration of high-achieving adolescents. While prior research has found an overall positive relationship between achievement and aspects of social integration, results have not been fully conclusive. First of all, it has been shown that the strength of the relation may differ based on the indicator used to measure achievement (Wentzel et al., 2021; Wolter & Seidel, 2017), the school subject under study (Wolter & Seidel, 2017) and the indicator used for measuring peer acceptance (Vörös et al., 2019; Wolter & Seidel, 2017). However, these variations have not yet been systematically assessed. Secondly, most studies so far, at least implicitly through their analysis protocols, assumed a linear relationship between achievement and social acceptance and did not compare different groups along the achievement distribution. Only a few studies that focused on stereotyping or bullying of high achievers have utilized non-linear relationships (Bergold et al., 2020). Thirdly, the role of gender for this relationship is not yet clear. Different findings regarding gender differences in the relation between achievement and social integration might be further understood when examining moderating factors like achievement domain.

In the present study, we will be examining the following hypotheses.

Hypothesis 1: There is an overall positive relationship between achievement and the sub facets of social integration on an individual level. To obtain a differentiated picture, different measures of achievement and social integration will be compared. However, we have no a-priori hypotheses regarding differences in the strength of relationships between the combinations of achievement and integration measures.

Hypothesis 2: However, this relationship might not be linear. Instead, we expect a curvilinear relationship with social integration decreasing for the highest levels of achievement. We do not

hypothesize about where on the achievement continuum this turning point might be. Therefore, the possible turning point is an exploratory question in the current analysis. We base the prediction of a non-linear relationship on studies suggesting that many students hold negative stereotypes of high-achievers (Boehnke, 2008; Pelkner & Boehnke, 2003; Workman & Heyder, 2020). Also, students in our study are around 15 years old, that is, they are within a developmental phase when school success as source of popularity becomes less important (Vannatta et al., 2009).

Hypothesis 3: This should be true especially for boys, that is, the drop of social acceptance in the highest achievement levels should be more pronounced for boys. This expectation follows from research showing that striving for school success is often considered to be a more feminine trait (Jones & Myhill, 2004). Thus, contrasting high-achieving girls and boys, boys seem to be viewed more unfavorably, generally (Bergold et al., 2020).

Hypothesis 4: The relationship is different with respect to the academic subjects under consideration. For mathematics and sciences, social integration of high achievers might be lower than for language arts. This prediction follows from the finding that students who are high in language competences are seen as more socially able, while the stereotype for scientists is that they are less social (Händel et al., 2013; Händel et al., 2018; Hannover & Kessels, 2004). Students who show the highest levels of achievement in all academic domains might show the most explicit drop in levels of integration, however, this is an exploratory hypothesis as prior research on this topic has not investigated profiles of achievement.

Hypothesis 5: There is an interaction between gender and subject, meaning that non-conforming high achievement in gender-stereotyped school subjects (e.g., boys performing exceptionally well in language arts, girls performing exceptionally well in mathematics and "hard" sciences) results in less social integration. It has been shown that male students show on average higher competences, higher self-concept of achievement and higher interest in mathematics, physics and chemistry and that German and biology are more female dominated subjects (Stanat et al., 2019b).

Methods

Data and Sample

We used data collected as part of the German National Trends in Student Achievement Study of 2018 (IQB-Bildungstrend 2018; Stanat et al., 2019b). The assessment had taken place in line with ethical regulations and was approved by the education ministries of the German federal states. The data comprises information on a nationally representative sample of about 45,000 9th grade students from circa 2,100 school classes, their parents, teachers and principals.

We have included students from regular schools in the highest, intermediate, and lowest tracks of the German tracking system. To construct our analysis dataset, special educational needs schools (N = 241 schools, N = 1,756 students) and 8 students at unknown school types have been excluded, as these schools constitute a specific developmental context which is not comparable to the other schools in the German tracked education system (Maaz et al., 2008). Next, we excluded N = 768 students with special educational needs at regular schools which did not follow the regular curriculum and have different grading standards. These steps left us with a sample of N = 42,370 students from 1,998 German 9th grade classrooms.

Variables and Instruments

Achievement

We pay special attention to the way achievement is operationalized. Especially, we vary the indicator used (grades vs. test-scores vs. self-concept), the domain focus (subject specific, mean scores vs. achievement profiles) and the comparison standard.

The dataset contains students' last report card grades in German, mathematics, and sciences. (In some schools, science is taught as a joint subject, in others, physics, chemistry and biology are separate courses with separate grades.) Students in Germany are graded on a scale from 1 – 6 with 1 being excellent, 2 good, 3 average, 4 sufficient and 5 and 6 failed. As the number of students receiving a grade 6 was very small (see Table III-1), we collapsed grades 5 and 6 into a common category.

Achievement test scores for mathematics and subject knowledge in physics, chemistry and biology were derived from students' responses to the German national competence assessment. These tests were constructed on the basis of the German curriculum and scaled using the framework of item response theory. In our analyses, we used plausible values present in the data set (15 for each domain). Details regarding the assessment and scaling procedure are given in the report (Stanat et al., 2019a, 2019b).

While objective achievement levels were our center of attention, prior research also suggested that the relationship between academic achievement and social integration might be mediated by academic self-beliefs. Thus, social integration and social interactions might be more closely related to the way high achieving students act. For this achievement-related social behavior, academic self-concept might be a critical explanatory variable, as also suggested by Wolter and Seidel (2017). We therefore contrasted high achievement with academic self-concepts of achievement to see which factor correlates more closely with the different facets of social integration. Self-

concept of achievement was assessed separately for the subjects mathematics, physics, chemistry and biology. Students indicated in how far four statements regarding their conception of their own ability (e. g. “I am a quick learner in mathematics.” or “I am good at mathematics.”) applied on a four-point-scale with 1 indicating complete rejection and 4 indicating full endorsement. Reliabilities for the four subjects were excellent (Cronbach’s $\alpha > .90$). All achievement variables have been coded so that greater values correspond with greater performance.

Social Integration

The sub-facets of social integration were assessed with psychometric scales for self-perceived social integration and school belonging as well as a sociometric questionnaire regarding friendship choices, interactions, helping behavior and rejection.

In the sociometric questionnaire, the students were presented with a roster which listed of all of their classmates. They then had to indicate 1) which of their classmates they would consider as friends, 2) who they usually spend their breaks with, 3) who they would ask for help if encountering problems and 4) who they would not like to sit next to. They had the opportunity to choose an unlimited number of their classmates. Classrooms in which less than half of the students participated in the network questionnaire ($N = 290$) were set to missing on all sociometric items, to exclude them from analyses of network measures. Data from the social network questionnaire were included for $N = 1708$ classes. The measures for the different sub-facets of social integration (friendship, contact, acceptance, and perceived acceptance) are described in the following.

Table III-1
Descriptives of Student Achievement

Variable	Mathematics	Physics	Chemistry	Biology	German
Grades					
1	3.6%	6.7%	6.6%	7.8%	7.6%
2	23.6%	22.6%	26.4%	26.6%	29.5%
3	44.3%	33.8%	37.4%	36.7%	37.8%
4	24.5%	27.4%	24.2%	23.6%	20.9%
5	3.8%	9.1%	5.1%	4.9%	3.8%
6	0.2%	0.5%	0.3%	0.4%	0.3%
Self-concept of achievement					
M(SD)	2.6 (0.8)	2.5 (0.7)	2.5 (0.8)	2.8 (0.7)	
Competence					
Optimal standard	3.6%	4.2%	2.4%	1.3%	
M(SD)	505.1 (93.4)	502.7 (92.5)	498.8 (94.2)	502.1 (94.1)	

Note. N_{total} = 42,370. Optimal standard is the highest achievement level (out of five) in the competence assessment used.

Friendship was assessed using the sociometric questionnaire item on which students indicated their friends. On this basis, we used two alternative ways to operationalize friendship: (a) number of reciprocal friends, normalized at classroom level to account for different class sizes and (b) having at least one friend.

Contact according to Bossaert et al. (2013) is implicated by participation in group activities, spending free time together and not being isolated. It was therefore assessed with sociometric item "Who do you usually spend your break with?" In order to get a measure which is independent of the subjective view, we used in-degrees, that is, normalized number of nominations received on this question. Second, the aspect of social isolation is considered by coding whether a person is completely isolated, that is, has no one spend their break with (zero in- and outdegrees).

Bossaert et al. (2013) subsume social preference, social support and social rejection under the term social acceptance. Although social preference would be indicated by the number of friendship nominations, we decided not to use this indicator, because the way we could measure it, it would be confounded with the friendship dimension of social integration. Instead, we used the sociometric question of whom students ask for social support if they have a problem. Students who are asked for help frequently (indegrees) are assumed to be socially accepted. Finally, social rejection was measured using indegrees for the fourth social network item: "Who don't you want to sit next to?". This indicator was recoded so that greater values denote greater integration.

Finally, self-perceived peer acceptance was also operationalized by two different indicators. First, subjective social integration was examined using a 7-item scale (for example "I have many friends in my class."; "I like to spend the break with others from my class."; "Others seem to like me.") which was to be rated by the students themselves. Answers were given on a scale between 1 (no endorsement) and 4 (full endorsement). The reliability was good (Cronbach's $\alpha = .86$). Second, school belonging, was assessed on a 9-item scale including statements like "I feel like an outsider at my school." "Other students seem to like me." "I feel lonely at my school." Again, the reliability was satisfactory (Cronbach's $\alpha = .88$).

Analyses

Regarding the *first research question*, we aimed to systematically evaluate the variability of the relationship (i.e., the bivariate correlation) between achievement and social integration across different operationalizations of achievement and between the different sub-facets of social integration. To this end, we conducted a multiverse analysis (Steegen et al., 2016) and depicted results as a specification curve (Simonsohn et al., 2020). Multiverse analysis (Steegen et al., 2016) aims to detect whether results of a study hinge on specific (to a certain degree arbitrary) decisions

in the process of data construction. First, a number of decisions during the process of creating an analysis data set from the raw data file are identified. Then, for each of these decisions, a set of defendable and sensible alternatives are listed. In this fashion, decisions (e. g. about the way in which important constructs are operationalized or which subjects are excluded from analysis) are systematically varied and all combinations are assembled to produce a multiverse of possible datasets. Next, the resulting multiverse of data sets is analyzed, that is, for each data set, an individual set of results is obtained. Finally, variability in results which is attributable to these decisions is made visible. The specification curve approach (Simonsohn et al., 2020) is a useful tool for visually inspecting the results of a multiverse analysis by juxtaposing effect sizes and their corresponding specifications. In our analysis we have varied the way achievement and social integration were operationalized. While we do not consider the different specifications as arbitrary variants of the same constructs, we wanted to leverage the proposed method to explore and compare the different constructs and their interrelations. Specifically, we varied the indicator used (grades, test scores, and academic self-concept), the subject and subject specificity of the achievement construct (mean scores vs. subject specific achievement in German, mathematics and science) and the reference standard chosen (population reference, classroom reference, or criterion reference). For each of the above-mentioned indicators of social integration, a multiverse of 84 combinations resulted (see Table III-2 for an overview of specifications).

The visual inspection of the variable distributions showed that the achievement variables could be considered normally distributed. However, the distribution of most of the social integration variables was truncated and therefore skewed. Therefore, we calculated spearman correlations for each combination of achievement and social integration. The correlations were made comparable by Fisher Z transformation.

For the *second hypothesis*, we first performed a two-lines test (Simonsohn, 2018), which tests for the existence of a (reversed) u-shaped curve in the relation between two variables. In addition, we estimated non-linear regression models predicting social integration by achievement. We entered a squared achievement term in our linear regression analyses and, for those variables in the network questionnaire which were truncated between 0 and 1, we also estimated quasi-binomial regressions.

Whereas our goal for the multiverse analysis was to systematically vary indicators for achievement and social integration, we conducted the regression models only for a narrower subset. That is, we focused on mean grade scores across all subjects as the achievement indicator. We centered the average grade score at a value of 2 in the original metric, because we expected to find non-linear

effects at the upper end of the achievement scale. Further, we recoded achievement for the analyses so that higher coefficients and a positive sign refer to higher achievement levels.

For the *third research question*, we calculated interactions between achievement and gender variables within the above-mentioned regression models to find out whether the effects of achievement on social integration differed for girls and boys.

For *Hypothesis 4*, we estimated regression models with achievement groups based on subject of achievement. For each student, a profile was coded, indicating whether a student was (a) a top-grade student in mathematics or physics only, (b) a top-grade student in English or biology only, (c) a top grade student in both domains or (d) having a top grade in neither domain. We then inspected interactions with gender to find out whether there were differences regarding subject-specific advantages or disadvantages for high-achieving students from either gender (*Hypothesis 5*).

In the regression analyses, we included a number of covariates. Firstly, we included classroom network density and size as covariates, because these have been shown to influence the possibility of nominating others (Faust, 2006). Secondly, we included school type. Different school tracks are associated with highly different learning and social environments, especially in Germany (Baumert et al., 2006). Also, we included covariates on an individual level, which might be associated with students' popularity or with their friendship choices. We therefore included students' gender, socio-economic background, and language spoken at home. Missing values were dealt with by multiple imputation technique. Weighting was used so that results are representative for the population of 9th graders in Germany.

Transparency and Openness

The hypotheses and planned analyses have been preregistered (https://osf.io/hu2n8/). The data will be made available at the research data center of the Institute for Educational Quality Improvement, where researchers will be able to use the datasets for scientific research projects upon application. All code for conducting analyses and reproducing the figures and tables is publicly available (https://osf.io/hu2n8/). Analyses have been conducted with the software R, Version 4.1.1 (R Core Team, 2021). We have mainly relied on the packages eatGADS (Becker, 2021) for data preparation and management, wCorr (Bailey & Emad, 2021) and specr (Masur & Scharkow, 2019) for the multiverse analysis, eatRep (Weirich et al., 2021) for regression analyses, flextable (Gohel, 2021) and ggplot2 (Wickham, 2016) for tables and figures and groundhog (Simonsohn & Gruson, 2021) for enhanced reproducibility. All attached packages are listed in Table III-3.

Results

Tables III-1 and III-4 show sample descriptive statistics of student background, achievement, and raw social integration indicators. The following results are organized by analysis approach and targeted hypotheses.

Table III-2

Indicators Used for Measuring Achievement and Social Integration

Facette	Indicator
Achievement	
Grades	
	Subject specific
	Raw grades
	High achievement binary indicator[a]
	Classroom standardized grades
	General
	Mean grades
	Cross-subject top grade count
	Classroom standardized mean grades
Self-concept	
	Subject specific
	Classroom standardized self-concept of achievement
	General
	Mean classroom standardized self-concept of achievement
Competence	
	Subject specific
	Raw scores
	Classroom standardized competence scores
	Optimal standard binary indicator[b]
	General
	Mean score
	Classroom standardized mean competence scores
Social Integration	
Friendship	
	Reciprocal friendship nominations
	Having a friend binary indicator
Contact	
	Spending breaks together in-degrees
	Social isolation during breaks
Acceptance	
	Being asked for help in-degrees
	Rejection in-degrees
Self-perceived acceptance	
	Subjective social integration
	School belonging

Note. [a] with and without classes containing more than one top-grade student. [b]Highest competence level of the national assessment of student achievement.

Multiverse Analysis

Hypothesis 1: Positive relationship between sub-facets of social integration and different achievement indicators

Figure III-1 shows results for the correlational analyses in the form of specification curves where the correlation coefficient for a given set of indicators is plotted in ascending order (Masur & Scharkow, 2019; Simonsohn et al., 2020). Each panel refers to a different aspect of social integration. The figures indicate that the correlation between achievement and social integration was positive for most combinations of achievement and social integration indicators. The values range between Fisher's z = -.04 and Fisher's z = .38. The range of values was highest for social acceptance (M = .14, SD = .09; Panel C) followed by subjective social integration (M = .06,

Table III-3

List of packages loaded during analysis.

Packages loaded during analysis	
BIFIEsurvey (BIFIE et al., 2019)	nlme (Pinheiro et al., 2021)
bit (Oehlschlägel & Ripley, 2020)	officer (Gohel, 2021b)
cowplot (Wilke, 2020)	**papaja** (Aust & Barth, 2020)
DescTools (Signorell, 2021)	progress (Csárdi & FitzJohn, 2019)
dplyr (Wickham et al., 2021)	**psych** (Revelle, 2021)
eatGADS (Becker, 2021)	**purrr** (Henry & Wickham, 2020)
eatRep (Weirich et al., 2021)	sandwich (Zeileis & Lumley, 2021)
flextable (Gohel, 2021a)	sp (E. J. Pebesma & Bivand, 2005; E. Pebesma & Bivand, 2021)
Formula (Zeileis & Croissant, 2020)	SparseM (Koenker, 2021)
ggplot2 (Wickham, 2016)	**specr** (Masur & Scharkow, 2019)
gridExtra (Auguie, 2017)	**standardize** (Eager, 2021)
groundhog (Simonsohn & Gruson, 2021)	**stringr** (Wickham, 2019)
lattice (Sarkar, 2021)	survey (Lumley, 2004, 2010, 2021)
lavaan (Rosseel, 2012; Rosseel et al., 2021)	survival (Therneau, 2021)
lme4 (Bates et al., 2021)	tibble (Müller & Wickham, 2021)
lmtest (Hothorn et al., 2020)	viridisLite (Garnier, 2021)
magrittr (Bache & Wickham, 2020)	**wCorr** (Bailey & Emad, 2021)
Matrix (Bates & Maechler, 2021)	zoo (Zeileis et al., 2021; Zeileis & Grothendieck, 2005)
mgcv (Wood, 2021)	
mice (van Buuren & Groothuis-Oudshoorn, 2011, 2021)	

Note. Bold print are those packages that were explicitly called in the analyses.

SD = .05; see Figure III-1 Panel D) and contact (*M* = .03, *SD* = .03; see Figure III-1, Panel B), and was lowest for friendship as a facet of social integration (*M* = .01, *SD* = .02; see Figure III-1, Panel A) and, all facets displayed values close to zero at the lower end of the distribution of Fisher's z-values. Within the facet of self-perceived acceptance (see Figure III-1, Panel D), school belonging (abbreviated "SB" in the figure) was more strongly related to achievement than subjective social integration in the classroom (average correlations: *M* = .09 vs. *M* = .04, $t(77.95) = 6.00$, $p < .001$). This points to the importance to consider peers in the same school, but outside of the own classroom as important actors for the social experience for high-achieving students. The aspect of social integration which was most strongly related to grades was *being asked for help by peers*, an indicator of acceptance ("AFH" in Panel C). It might be the case that higher achievers are more strongly involved in instrumental relationships than in affective ones.

Turning to the different operationalizations of achievement, a number of points are apparent. In general, relations between integration and achievement were highest when competences were used as an indicator of achievement. Academic self-concept as a subjective indicator of achievement had rather low (and, in the case of friendship, even negative) associations with social integration. An exception regarding this general pattern was self-perceived acceptance, where higher levels of self-concept of achievement were related more strongly to subjective acceptance. This might reflect a general tendency of self-enhancement of some students.

As for the reference norm ("Ref" in Figure III-1), it is striking that using only those classes in which there was only one student being awarded the top grade (Grade 1, denoted "Top-student" in the figure), the correlations were quite small and in a number of cases (especially regarding friendship) even negative. Overall, the strongest correlations were found when using raw scores of grades and competence indicators of achievement; followed by classroom centered achievement scores, binary indicators (Grade 1 or highest proficiency level vs. all others) and—as mentioned before—only single top students in every class.

Another aspect is worthy of notice: For the sub-facet of friendship, achievement in German is more strongly related to social integration than all other subjects, whereas in all other sub-facets, a general achievement level seems more important.

Overall, thus, we find evidence for Hypothesis 1. For all four facets of social integration, the mean correlation across all indicators and domains was positive and the correlations were significant in the majority of scenarios.

Table III-4

Sample Descriptives

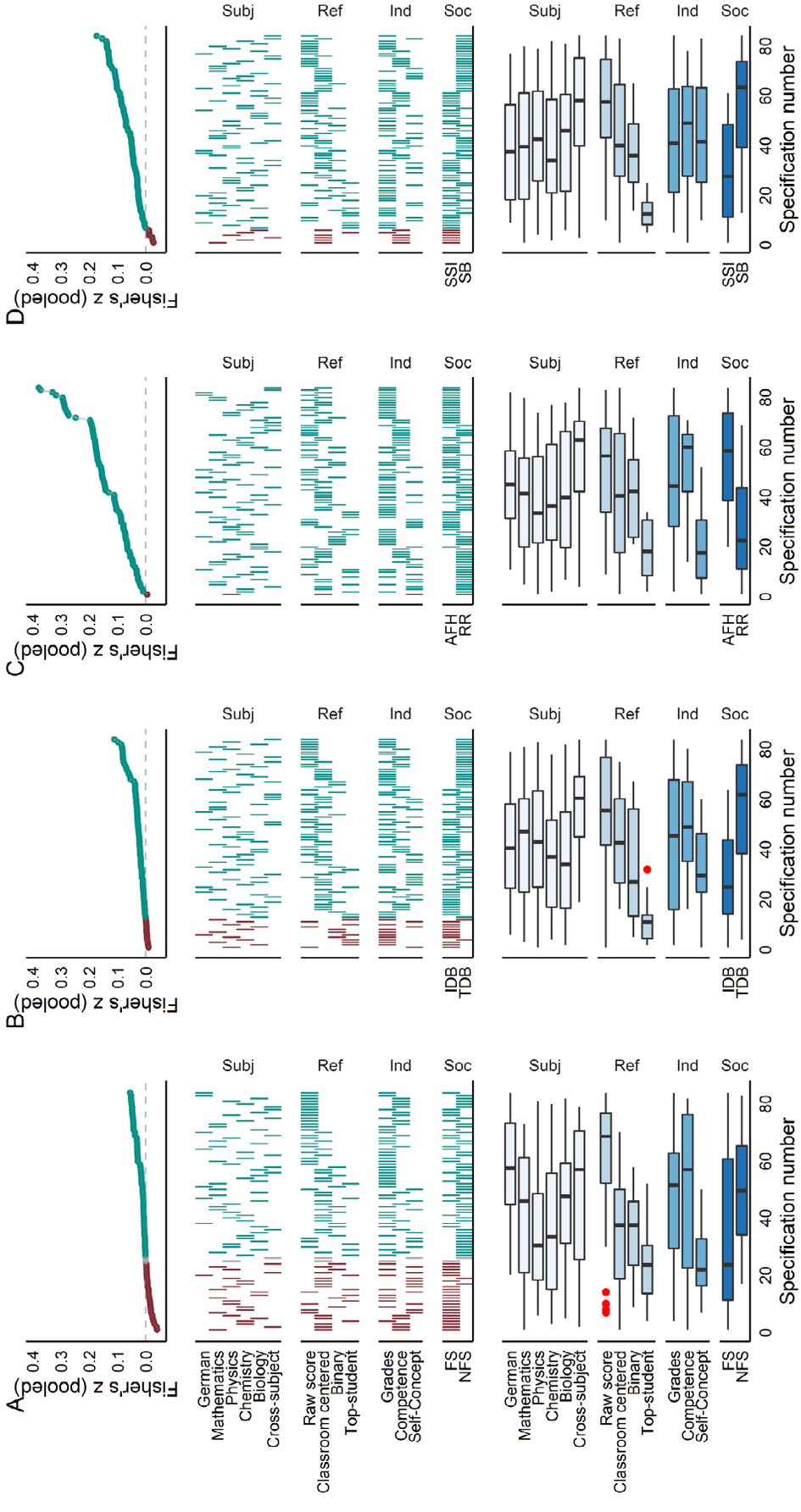
Variable	%	*M*	*(SD)*
Background characteristics			
Gender (girls)	48.8		
Age		15.57	0.64
Language at home (German)	69.4		
HISEI		51.26	20.51
School type: High track	36.8		
School type: Intermediate tracks	51.9		
School type: Low track	11.3		
Social integration			
Reciprocal friendship	26.7		
In-degree spending breaks	21.1		
In-degree help	20.1		
In-degree rejection	18.4		
Perceived social integration		3.25	0.51
School belonging		3.16	0.52

Note. N_{total} = 42,370. Social network parameters are normalized at classroom level and represent the number of received nominations for a student out of the total nominations possible in their classroom.

To make the following analyses, presentation, and discussion of results more concise, we narrowed our analysis down to one indicator of achievement: mean grades. We made this decision because it is most consistent with our focus on social processes in the classroom. Grades are a visible measure of student academic achievement and best reflect the social construction of the image of high achievers. Moreover, it can also be argued that grades are by definition the achievement outcome targeted by school education.

Figure III-1

Specification Curves for the Correlation Between Achievement and Facets of Social Integration



Note. For each facet of social integration, a specification curve is depicted which shows the correlation between achievement and social integration. The top panel shows the distribution of Fisher's Z values across the combinations of achievement and social integration indicator (i.e., the specifications). The middle panel indicates the specification belonging to the respective value in the top panel. Achievement indicators are characterized by a unique configuration of the subject of achievement ("Subj"), the reference ("Ref") and the indicator ("Ind") chosen. Social integration indicators ("Soc") are explained below. The bottom panel depicts the marginal distribution of Fisher's Z by each specification factor. Panel A: Friendship. NFS = No reciprocal friendship nominations (recoded), FS = Reciprocal friendship nominations Panel B: Contacts. TDB = Spending time during breaks together in-degrees, IDB = Social isolation during breaks (recoded). Panel C: Peer acceptance. RR = Rejection nominations (recoded), AFH = Being asked for help nominations. Panel D: Self perceived acceptance. SSI = Self-perceived social integration, SB = School belonging.

Two-lines test

Hypothesis 2: Non-linear relationship

As a first test of our non-linearity hypothesis, we conducted the two-lines test proposed by Simonsohn (2018). A u-shape (or an inverted u-shape) is supported if the algorithm identifies a cut-point at which a sign change between two linear regressions fitted to the data takes place. Table III-5 shows the results of this analysis for each of the social integration variables.

For reciprocal friendships, a solution detecting a u-shape was found, that is, both slopes were significantly different from zero and a cut-point within the range of data was identified. The first slope was positive and the second one negative, indicating an inverted u. The cut-point was at 2.45 on the original grading scale, which corresponds to a good performance. That is, up until a mean

Table III-5

Results of Simonsohn's (2018) two-lines tests for detecting an inverted u-shape

Dimensions	Slope 1			Slope 2			Cutpoint
	b_1	t	p	b_2	t	p	x_{orig}[a]
Reciprocal friendship nominations	0.02	10.49	***	-0.01	-2.31	*	2.45
Acceptance nominations	0.02	15.31	***	0.00	0.40		1.93
Being asked for help nominations	0.04	26.40	***	0.09	36.66	***	2.80
Rejection nominations (recoded)	0.04	37.85	***	-0.02	-0.71		1.40
Self-perceived social integration	0.07	16.44	***	-0.03	-0.80		1.73
School belonging	0.12	28.10	***	-0.05	-1.05		1.68

Note. [a]x_{orig} is the cutpoint transformed to the original grading scale of 1 – 5, with 1 being the top grade and 5 representing the failing grades.
* p < .05, *** p < .001.

grade of 2.45, students' friendship ties increased when their grades increased. Beyond that point, better grades were not beneficial, but slightly detrimental to social integration. For the amount of being asked for help, a cut-point at grade 2.80 (which corresponds to a rather mediocre level of achievement) was found, separating a modestly positive segment from a more pronounced positive regression line. This pattern is not indicative of a u- or inverted u-shape but rather of a non-linear accelerated positive trend. None of the other indicators of social integration showed a u-shape in this first test. For all of them, the first slope had a positive sign, but the second slope did not significantly (at an α-level of .05) differ from zero, indicating a plateau. An example plot from the two lines test is given in Figure III-2.

Hypothesis 3: Gender differences in the relationship between achievement – social integration

As a next step, we conducted the two-lines test separately for girls and for boys. Comparative results for these groups are given in Table III-A1 in Appendix III-A. Here, the two lines tests were not significant which may be due to power constraints. The boys, however, exhibited a more pronounced non-linear relationship between achievement and being asked for help. Their first slope is flatter and the second slope is steeper than the girls' slopes. Consequently, the cut-point maximizing the power to detect a u-shape is much lower for boys ($x_{orig} = 2.8$) than for girls ($x_{orig} = 1.4$).

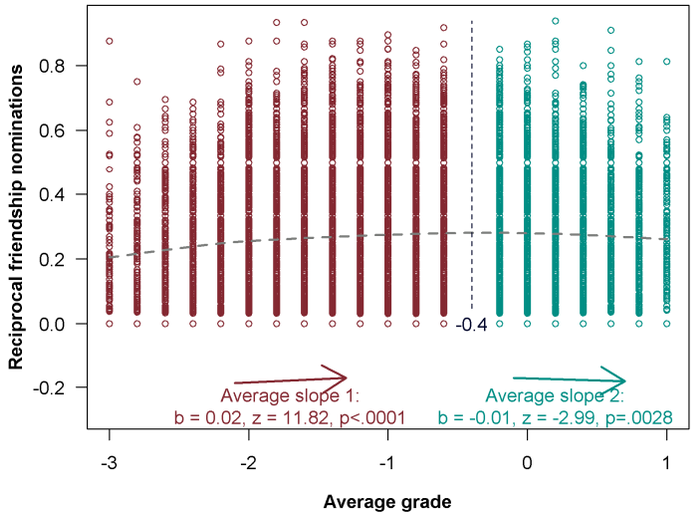


Figure III-2. Result from Two Lines Test of the Relationship Between Grades and Reciprocal Friendship Nominations. (This plot shows results from the two-lines test using one imputation of the data. The plot was created based on the syntax of the Two-line test app (Simonsohn, 2018: http://webstimate.org/twolines/).)

Regression models

Table III-6 summarizes the results of different linear and non-linear (quadratic, quasi-binomial) regression models linking achievement in terms of average school grades and different facets of social integration. Tables III-7 and III-8 demonstrate the statistical analyses used and detailed results for the facet of friendship. The same model was applied to each of the other facets, that is, to contact (spending breaks together), acceptance (being asked for help and rejection nominations), and subjective acceptance (contact out-degrees, subjective social integration and school belonging). These individual regression tables can be found in Appendix B.

Hypotheses 1 and 2: Positive, non-linear relationship between achievement and social integration

It is evident that, confirming the results of the earlier correlation analyses, the general positive main effect of achievement on different indicators of social integration is maintained after controlling for covariates (H1 sustained).

Table III-6

Summary of Results for Multiple Regression Analyses

Social Integration	ME ACH	QD ACH	ME GEN	IA ACH*GEN	ME DOM: none/MP/EB	IA DOM*GEN
Friendship						
Reciprocal friendship nominations	+	-	0	0	0/0/0	0/0/0
Contact						
Contact nominations	+	-	-	0	-/0/0	0/0/0
Acceptance						
Being asked for help nominations	+	+	+	-	-/-/-	+[a]/0/0
Rejection nominations (recoded)	+	-	+	-[b]	-/0/0	0/0/0
Subjective social integration						
Self-perceived social integration	+	0	0	0	-/0/0	0/0/0
School belonging	+	0	-	0	-/-/0	0/0/0

Note. The table summarizes the results of nine or five individual regressions for each sociometric or psychometric indicator of social integration (rows), respectively. ME = main effect, QD = quadratic term, IA = Interaction, ACH = achievement, GEN = gender, DOM = domain of achievement, MP = mathematics/physics, EB = English/biology. + indicates a significant positive effect, - indicates a significant negative effect, 0 indicates non-significant parameter estimates. [a]Effect is significant only in the quasibinomial model parametrization. [b]Interaction is not significant in the quasibinomial model.

Regarding non-linear effects, the picture is differentiated. For subjective social integration, no non-linear effect was found. For the dimension of helping behavior, a positive quadratic effect emerged, that is, more highly achieving students were sought out for help more often, a result consistent with the results of Simonsohn's two-lines test reported above. For all other dimensions of social integration, a negative quadratic effect occurred. As can be seen from Figure III-3, the model-implied relationship between reciprocal friendship nominations and average grade has its maximum within the range of the achievement scale, thus leading to a reduction in the number of reciprocal friends for the highest achievers. This result mirrors the findings from Simonsohn's u-test. For acceptance and rejection, the relationship between achievement and social integration decreases towards the higher achievement levels, but there is no negative turn within the range of the grading scale. Comparing the different models, although the quadratic term turned significant in these models, the improvement in R^2 was only marginal.

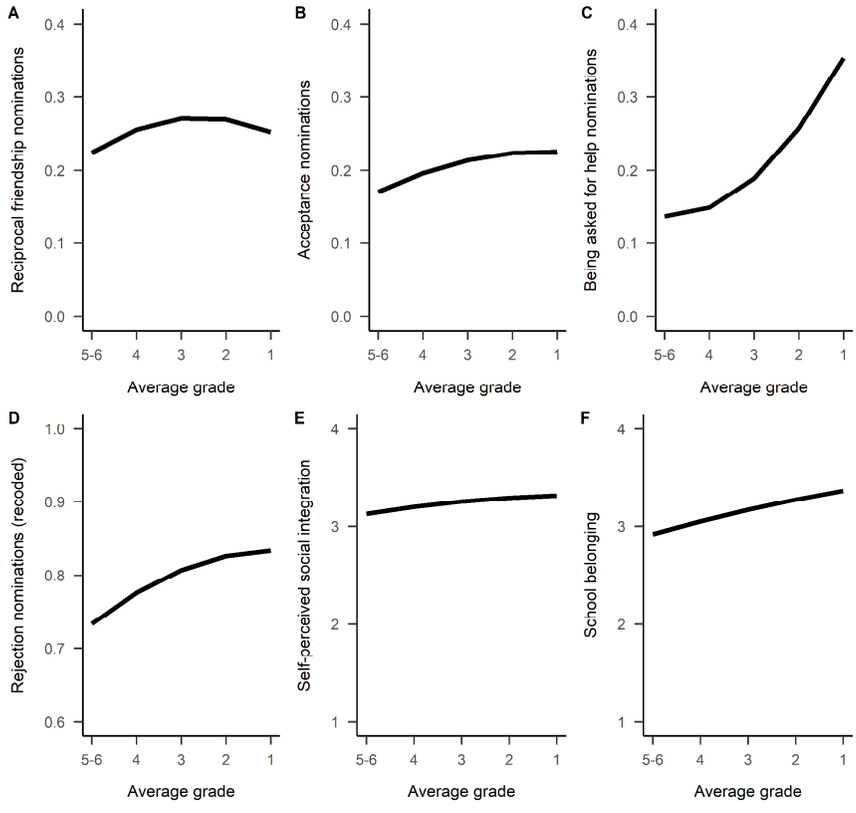


Figure III-3. Quadratic Models of the Relationship Between Average Grade and Social Integration. (Plots visualize the predicted values from the quadratic models (i.e., Model (2) in Table III-7). Quadratic terms were significant for all models including sociometric indicators (Panels A - D).)

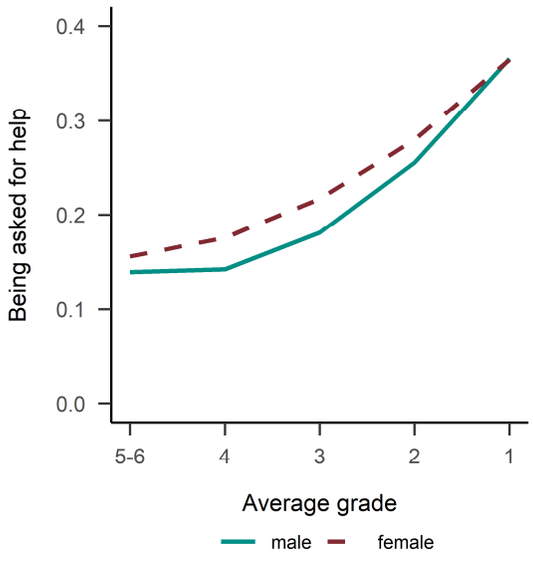


Figure III-4. Interaction Effects Between Achievement and Gender for Nominations of Being Asked for Help. The plot shows the estimated number of peers (normalized at classroom level) who would ask this student for help. The estimates are derived from the quadratic model with interaction effect. Covariates were fixed at the sample means.

Hypothesis 3: Gender differences

To test whether the nonlinear shape was different for girls and boys, we next conducted regressions including interaction terms between achievement and gender. However, the only dimension of social integration for which a gender interaction occurred was the social network indicator of being asked for help (see Tables III-6 and III-B3). As Figure III-4 reveals, girls are generally more sought after when it comes to help and advice. Only the most highly achieving boys were equally asked for help.

Hypothesis 4: Subject differences

To test our hypothesis regarding differences between academic subjects, we used dummy variables denoting if a student was (a) highly achieving in mathematics or physics (typically male stereotyped subjects), (b) highly achieving in German or Biology (female stereotyped subjects), (c) highly achieving in both domains or (d) highly achieving in none of the domains. There were effects of the achievement profile regarding the sociometric help-seeking question. Students who are high achieving in both domains are asked for help more often than students who are highly achieving in none or in only one of the domains. For the friendship dimension, there were no group differences regarding the subjects of achievement, and for the dimensions acceptance, rejection and subjective integration (self-perceived integration, school belonging and contact out-degrees), there were negative effects for students not high achieving in any of the domains, but between the different subject profiles, there were no differences in integration (see Table III-6, for a summary).

Table III-7

Regression of Reciprocal Friendship Nominations on Achievement and Gender

Variable	(1)			(2)			(3)		
	B	*p*	β	*B*	*p*	β	*B*	*p*	β
Intercept	0.268	***		0.267	***		0.263	***	
Average Grade (linear term)	0.005	***	0.032	-0.009	**	-0.058	-0.013	**	-0.083
Average Grade (quadratic term)				-0.008	***	-0.051	-0.008	***	-0.051
Gender (girls)	0.001		0.006	0.001		0.006	0.007		0.045
Average Grade *Gender							0.005		0.032
Average Grade (quadratic term) *Gender							-0.001		-0.006
Covariates									
Network density	0.836	***	5.356	0.836	***	5.356	0.835	***	5.349
Class size	-0.001		-0.006	-0.001		-0.006	-0.001		-0.006
School type (highest track)	0.004		0.026	0.004		0.026	0.004		0.026
School type (lowest track)	-0.014	**	-0.090	-0.014	**	-0.090	-0.014	**	-0.090
HISEI	0.000	**	0.000	0.000	*	0.000	0.000	*	0.000
Language	0.002		0.013	0.001		0.006	0.001		0.006
R-squared	0.292			0.295			0.295		

Note. β are regression parameters standardized by y, that is, the parameters denote the change in standard deviations of friendship nominations at a one-unit increase in the predictor. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992). The strong effect of the covariate network density occurs because the level of interrelatedness of students in a classroom naturally influences the possibility of receiving nominations. Faust (2006) has pointed out this fact and recommended controlling for its effect.
* $p < .05$, ** $p < .01$, *** $p < .001$

Hypothesis 5: Gender-subject interactions

Because we had the hypothesis that achievement in opposite gender typical subjects would result in lower social integration, we further inspected the interaction effects between gender and achievement domain. We did not find any of the hypothesized interaction effects. Only one interaction effect (related to being asked for help) became significant (see Table III-B6 in Appendix III-B). Girls who were not achieving highly in any of the four subject domains were asked more often for help than boys. However, the effect was inconsistent and occurred only in the quasibinomial model.

Discussion

Our study examined the relation between academic achievement and social integration using a variety of indicators. Overall, we could confirm earlier findings (Wentzel we al., 2021) of a generally positive (albeit weak) relationship between social integration and achievement and found it across different operationalizations of achievement and different facets of social integration (Hypothesis 1). Inspecting the specification curves (Figure III-1) also showed that the indicators of social integration were quite differently related to academic achievement. Not only were there differences in the correlation with achievement between the four facets of friendship, contact, acceptance and subjective integration, but also for the different indicators within each facet. For example, within the facet of subjective social integration, school belonging was more highly correlated with achievement than self-perceived integration. Also, within the category acceptance, being asked for help and not being rejected as a desk neighbor are quite differently related to achievement, with generally higher correlations for being asked for help (up to Fisher's $Z = .35$).

Table III-8

Reciprocal Friendship Nominations Regressed on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)		
	B	*p*	*β*	*B*	*p*	*β*
Intercept	0.253	***		0.248	***	
Top grades in both domains (reference group)						
Top grades in neither domain	0.008		0.051	0.014		0.090
Top grades in German/Biology	0.010		0.064	0.021		0.135
Top grades in Math/Physics	0.005		0.032	0.005		0.032
Gender (girls)	0.002		0.013	0.011		0.070
Top grades in neither domain*Gender				-0.010		-0.064
Top grades in German/Biology*Gender				-0.017		-0.109
Top grades in Math/Physics*Gender				0.002		0.013
Covariates						
Network density	0.837	***	5.362	0.837	***	5.362
Class size	0.000		0.000	0.000		0.000
School type (highest track)	0.006	*	0.038	0.006	*	0.038
School type (lowest track)	-0.014	**	-0.090	-0.014	**	-0.090
HISEI	0.000	*	0.000	0.000	*	0.000
Language	0.003		0.019	0.003		0.019
R-squared	0.292		1.871	0.292		1.871

Note. β has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor. Average grade has been centered at 4 (equivalent to German Grade 2 out of 6).
* p < .05, ** p < .01, *** p < .001

This underscores the need to play close attention to the operationalization of social integration or social acceptance of students.

From the multiverse of results, one consistent pattern was that the highest correlations with achievement were found for being sought out for help. The finding that students who are achieving highly are asked for help more often than other students, especially if they perform well in multiple subjects, is a result that might be expected, if this help-seeking was about help in academic domains. Although the item formulation did not specifically target academic support, but asked for help-seeking in general, it could be argued that within the school context, where this study took place, academic help might be the first domain students would think of. On the other hand, it could be possible that higher achievers might also have better socio-emotional and practical skills and are therefore asked for support in other matters, too. Differently from the other indicators of social integration, the highest correlations occurred if grades were used as a measure of achievement, followed by competences. It has been found that students who attain high grades are not only competent in the academic domain, but also possess high interpersonal competence (Wentzel, 1991). She discussed that, on the one hand, socially responsible behavior might contribute to grades by being a valued educational outcome of its own right. On the other hand, educational processes and socially responsible behavior might be causally linked in that learning processes engender social behavior or social behavior being a precondition for effective learning. However, relations between social, emotional or interpersonal intelligence (as a reason for being asked for advice) and abstract intelligence (which might be more closely related to test scores) are less clear (Kihlstrom & Cantor, 2011; Meijs et al., 2010). The pattern of a positive link between academic achievement and being asked for help also showed across different models that were used in the subsequent regression analyses.

Generally, the facet of friendship showed the weakest associations with achievement, and even a number of negative ones. These negative associations occurred mostly when academic self-concept was used as an indicator of achievement, but also if a classroom centered view was chosen and the singled out top student in a class was selected as high achiever. First of all, these low correlations indicate, of course, many factors other than academic achievement contribute to the desirability of a student as a friend. For example, one might think of extraversion or agreeableness (Harris & Vazire, 2016; Selfhout et al., 2010). Also, students for whom high academic performance is an important goal and an integrated part of their self-concept and who invest their time and resources into academic effort might place comparatively less value and effort in maintaining friendships. Wentzel and Asher (1995) have found that children who are sociometrically neglected and therefore do not have many friendship nominations, but at the same

time no negative ones, show generally very positive academic adjustment and better relationships with teachers. Therefore, a smaller number of friends might not be an indicator of poorer adjustment or a predictor of the feeling of social integration (Ferguson & Ryan, 2019; Gifford-Smith & Brownell, 2003).

Going beyond simple linear relationships, it was our endeavor to take a closer look at the shape of the association between academic achievement and social integration indicators. We hypothesized that there might be a non-linear trend, with social integration decreasing for the highest achievement levels (Hypothesis 2). The two-lines tests we conducted, indicated this inverted u-shape for reciprocal friendship nominations but not for the other facets of social integration. However, quadratic regressions showed a trend of attenuated social integration for all of the dimensions except for subjective social integration and help-giving. For help-giving, a cumulative effect of achievement occurred, that is, the higher the achievement level of a student, the more sought-after they were. Therefore, our second hypothesis was partly confirmed—on average across our modeling approaches, students with low achievement showed lower social integration than students with higher achievement, but at some point across the achievement distribution, higher achievement was not related to a further increase in social integration.

Hypothesis 3 stated that the proposed non-linear relationship should be especially pronounced for boys (i.e., boys paying a “higher price” socially for being in the highest achievement level). Our analyses disconfirmed this hypothesis. First, the interaction terms between gender and achievement in our regression models were not significant, with the exception of the integration indicator of being asked for help by peers. Here, boys had to be much higher achieving than girls to be asked for help as often. Because we had expected gender effects, we additionally conducted a further robustness check and examined the specification curve for boys and girls separately. Conducting the specification curve of the simple correlation between achievement and social integration separately for boys and girls, we found generally lower correlations for boys, but also in some cases stronger negative correlations (see Appendix A). This points to the possibility that high-achieving boys—compared to lower achievers— are generally seen in a less positive light than high-achieving girls. But also, variations in network nominations might depend on different predictors for boys and girls. If for boys, other aspects are more important in determining their popularity (as is suggested by Vannatta et al., 2009), relationships between popularity and academic achievement should be lower.

Our fourth hypothesis was concerned with possible differences between achievement domains. Our regression analyses revealed no consistently significant differences between male and female

stereotyped subject areas and facets of social integration. However, evaluating the specification curves, high achievement in German seems to be more strongly related to the facet of friendship than the other subject domains. In contrast, German achievement, across the specifications, seems less strongly associated with subjective social integration. For acceptance, and specifically the indicator of being asked for help, the association is strongest when looking at mean achievement across domains, an observation supported by the regression analyses in which being high-achieving in both subject areas is associated with being asked for help by significantly more students than being high-achieving in either one of the domains or in no domain. Thus, our hypothesis receives only weak support by our analyses, and it does so only in the facet of friendship. Also, Hypothesis 5, which asserted interactions between gender and subject of achievement, such that students performing high in opposite-gender stereotyped subjects would be chastised by socially, was disconfirmed in our analyses. Most studies on which we rested this assumption worked with specifically asking students about the stereotypes they held, for example by demanding them to rate fictious new high achieving classmates or by asking about fears students experience regarding high achievement (Kessels, 2005; Smith & Leaper, 2006). Thus, there seems to be an inconsistency between what students expect to be happening to high achievers (stereotypes) and what does really happen to high achievers in the social context of the classroom (sociometric popularity), generally.

Limitations

Our study suffers from several limitations, which we would like to address. Firstly, the dataset that was used did not include other variables which might predict social integration, like personality (Rentzsch et al., 2013). Including such variables as controls might have permitted to paint an even clearer picture of the relationship between achievement and social integration. For example, certain personality traits (like conscientiousness or agreeableness) might influence both academic achievement and likeability by peers, thus confounding the relationship between the two constructs. Likewise, gender differences in personality traits have been found (Costa et al., 2001), which might obscure the relationship between achievement, gender and social integration. Thus, future research could shed additional light on the relationship between social integration and achievement, by including these different variables in the data collection. Another variable which was not available was academic self-concept and competence in German, as the assessment focus in 2018 was mathematics and science. Therefore, while we had grades in all achievement domains, the other achievement indicators refer to mathematics and sciences, only. This could be making a difference in the interpretation of the specification curves which should be kept in mind.

Another point which might be raised is the fact that this cross-sectional design doesn't allow for causal inferences. The relationships between these constructs—social integration and academic achievement—are probably complex and bi-directional in nature. Earlier research has mainly argued for mechanisms by which social behavior and social integration precedes academic achievement, in the form of grades (Shin & Ryan, 2014; Wentzel, 1991). However, sociometric research on friendship formation also finds evidence for selection processes, that is, students select to spend time with peers based on their academic attainment (Gremmen et al., 2017). Additionally, the domain of help-seeking demonstrates that it is reasonable that some interpersonal relationships develop as a result of high achievement.

Finally, the use of regression models rests on the assumption that observations are independent of each other. However, because relationships exist between individual students, social integration scores are dependent on the scores of other children in the same class. To incorporate this dependency in our analyses, we firstly used normalized nomination scores at class level, so that every student's scores can vary between 0 and 1. Also, we included covariates related to properties of classroom networks, like density, class size and school type. One alternative class of models that deals with this dependency are the exponential random graph models for social networks (Robins et al., 2007). These use network ties as units of analysis and assess the probability of a tie between two actors in a specific network. Thus, they would be applicable for those social integration indicators using network ties. However, these models have the drawback that every classroom must be modeled separately in a very cumbersome manner and are thus resource intensive. In a future project we aim to confirm and extend the findings from the present study.

Conclusion and Outlook

The take-home message of this study is a positive one: Overall, high-achieving students seem not to be at high risk for social exclusion. Rather, the higher the achievement level, the more positive are the social integration indicators are in general even though this relation becomes weaker with higher achievement (i.e., achievement seems to have diminishing returns with regard to social integration). Only the number of friends seems to decrease a little for the topmost levels of achievement, a finding which might point to the relative importance different students attach to academic vs. social lives. In how far the number of reciprocated friends a student has relate to differences in satisfaction with social relationships between high- and lower achievers, remains an open question. Moreover, often stated fears which reflect in stereotypes about boys and girls of high performance in different academic subjects (Boehnke, 2008; Masters et al., 2021; Workman & Heyder, 2020) seem not to be warranted, at least from this research that incorporated a wide range of indicators for social integration.

The study was conducted with a representative sample in Germany. Therefore, the generalizability to students in other countries is open for debate. Countries differ in their school systems with regard to the heterogeneity within classes, with regard to instruction and organization of schools and with regard to cultural values. Some comparative research has found considerable differences in stereotypes of high-achievers and their social standing in different cultures (Händel et al., 2018; Wentzel et al., 2021). Therefore, investigating the relations between actual social integration and achievement and its shape across different countries seems to be a promising avenue for further research.

Likewise, the construct of social integration deserves more attention by research and theorizing. Already, some research has tried to determine relationships between different concepts like acceptance, friendship and victimization (Ladd et al., 1997), or sociometric status, friendship and network position (Gest et al., 2001; Gifford-Smith & Brownell, 2003). Also, some research has been concerned with measurement issues, both conceptually and empirically (Vörös et al., 2019; Wentzel et al., 2021). However, bringing together the sub-facets of social integration identified by Bossaert et al. (2013) and Koster et al. (2009) using conceptual analysis with empirical applications, relying on both self-reports of integration as well as sociometric measures is a unique contribution of this piece of research. It might be seen as a starting point to further differentiate the concept of social integration and for determining whether the identified facets can be operationalized and discriminated empirically. In the following, research could determine how these different indicators and facets of social integration can be predicted by different person or environment variables and how they relate to other outcomes, such as well-being or academic adjustment. Thus, the next questions that arise are how are different aspects of social integration related to each other, how do they predict psychosocial and academic outcomes and how can they be influenced?

References

Abrams, D., Rutland, A., & Cameron, L. (2003). The development of subjective group dynamics: Children‘s judgments of normative and deviant in-group and out-group individuals. *Child Development*, *74*(6), 1840–1856. https://doi.org/10.1046/j.1467-8624.2003.00641.x

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2013). ‘Not girly, not sexy, not glamorous’: primary school girls’ and parents’ constructions of science aspirations. *Pedagogy, Culture & Society*, *21*(1), 171–194. https://doi.org/10.1080/14681366.2012.748676

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for „Grid“ Graphics*. https://CRAN.R-project.org/package=gridExtra

Aust, F. & Barth, M. (2020*). papaja: Prepare reproducible APA journal articles with R Markdown*. R package version 0.1.0.9997, retrieved from https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2020). *magrittr: A Forward-Pipe Operator for R*. https://CRAN.R-project.org/package=magrittr

Bailey, P., & Emad, A. (2021). *wCorr: Weighted Correlations*. https://american-institutes-for-research.github.io/wCorr/

Bates, D., & Maechler, M. (2021). *Matrix: Sparse and Dense Matrix Classes and Methods*. http://Matrix.R-forge.R-project.org/

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2021). *lme4: Linear Mixed-Effects Models using Eigen and S4*. https://github.com/lme4/lme4/

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 95–188). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90082-7_4

Becherer, J., Köller, O., & Zimmermann, F. (2017). Sozialverhalten und Schulleistungen: Spielt die Beliebtheit in der Klasse eine Rolle? [Social behavior and school achievement. Does peer acceptance play a role?]. *Zeitschrift für Erziehungswissenschaft*, *20*(3), 405–424. https://doi.org/10.1007/s11618-017-0771-1

Becker, B. (2021). *eatGADS: Data Management of Large Hierarchical Data*. https://github.com/beckerbenj/eatGADS

Benenson, J. F. (1990). Gender Differences in Social Networks. *The Journal of Early Adolescence*, *10*(4), 472–495. https://doi.org/10.1177/0272431690104004

Bergold, S., Kasper, D., Wendt, H., & Steinmayr, R. (2020). Being bullied at school: the case of high-achieving boys. *Social Psychology of Education*, *23*(2), 315–338. https://doi.org/10.1007/s11218-019-09539-w

BIFIE, Robitzsch, A., & Oberwimmer, K. (2019). *BIFIEsurvey: Tools for Survey Statistics in Educational Assessment*. https://CRAN.R-project.org/package=BIFIEsurvey

Boehnke, K [Klaus] (2008). Peer pressure: a cause of scholastic underachievement? A cross-cultural study of mathematical achievement among German, Canadian, and Israeli middle school students. *Social Psychology of Education*, *11*(2), 149–160. https://doi.org/10.1007/s11218-007-9041-z

Bond, R. M., Chykina, V., & Jones, J. J. (2017). Social network effects on academic achievement. *The Social Science Journal*, *54*(4), 438–449. https://doi.org/10.1016/j.soscij.2017.06.001

Boor-Klip, H. J., Segers, E., Hendrickx, M. M. H. G., & Cillessen, A. H. N. (2017). The Moderating Role of Classroom Descriptive Norms in the Association of Student Behavior With Social Preference and Popularity. *The Journal of Early Adolescence*, *37*(3), 387–413. https://doi.org/10.1177/0272431615609158

Bossaert, G., Colpin, H., Pijl, S. J., & Petry, K. (2013). Truly included? A literature study focusing on the social dimension of inclusion in education. *International Journal of Inclusive Education*, *17*(1), 60–79. https://doi.org/10.1080/13603116.2011.580464

Brown, B. B., & Larson, J. (2009). Peer relationships in adolescence. In R. M. Lerner & L. D. Steinberg (Eds.), *Handbook of adolescent psychology: Contextual influences on adolescent development* (pp. 74–103). John Wiley & Sons.

Cairns, R., Xie, H., & Leung, M. C. (1998). The popularity of friendship and the neglect of social networks: Toward a new balance. *New Directions for Child Development*(80), 25–53. https://doi.org/10.1002/cd.23219988104

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Csárdi, G., & FitzJohn, R. (2019). *progress: Terminal Progress Bars*. https://github.com/r-lib/progress#readme

De Laet, S., Doumen, S., Vervoort, E., Colpin, H., van Leeuwen, K., Goossens, L., & Verschueren, K. (2014). Transactional links between teacher-child relationship quality and perceived versus sociometric popularity: A three-wave longitudinal study. *Child Development*, *85*(4), 1647–1662. https://doi.org/10.1111/cdev.12216

Eager, C. D. (2021). standardize: Tools for Standardizing Variables for Regression in R. https://github.com/CDEager/standardize

Faust, K. (2006). Comparing social networks: size, density, and local structure. *Metodoloski Zvezki*, *3*(2), 185–216. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.8247&rep=rep1&type=pdf

Feiring, C., & Lewis, M. (1991). The Development of Social Networks from Early to Middle Childhood: Gender Differences and the Relation to School Competence. *Sex Roles*, *25*(3), 237. https://search.proquest.com/scholarly-journals/development-social-networks-early-middle/docview/1308098505/se-2?accountid=11531

Ferguson, S. M., & Ryan, A. M. (2019). It‘s Lonely at the Top: Adolescent Students‘ Peer-perceived Popularity and Self-perceived Social Contentment. *Journal of Youth and Adolescence*, *48*(2), 341–358. https://doi.org/10.1007/s10964-018-0970-y

Festinger, L. (1952). Informal social communication. In L. Festinger, K. Back, S. Schachter, H. H. Kelley, & J. Thibaut (Eds.), *Theory and experiment in social communication* (pp. 1–18). Edwards Brothers, Inc.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics —A “hopeless” issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*(4), 497–514. https://doi.org/10.1007/BF03173468

Freund-Braier, I. (2009). Persönlichkeit [Personality]. In D. H. Rost (Ed.), Pädagogische Psychologie und Entwicklungspsychologie: Vol. 72. Hochbegabte und hochleistende Jugendliche: Befunde aus dem Marburger Hochbegabtenprojekt (2nd ed.). Waxmann.

Garnier, S. (2021). *viridisLite: Colorblind-Friendly Color Maps (Lite Version)*. https://CRAN.R-project.org/package=viridisLite

Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663–677. https://doi.org/10.1037/edu0000003

Gest, S. D., Graham-Bermann, S. A., & Hartup, W. W. (2001). Peer Experience: Common and Unique Features of Number of Friendships, Social Network Centrality, and Sociometric Status. *Social Development*, *10*(1), 23–40. https://doi.org/10.1111/1467-9507.00146

Gifford-Smith, M. E., & Brownell, C. A. (2003). Childhood peer relationships: social acceptance, friendships, and peer networks. *Journal of School Psychology*, *41*(4), 235–284. https://doi.org/10.1016/S0022-4405(03)00048-7

Gohel, D. (2021a). *flextable: Functions for Tabular Reporting*. https://CRAN.R-project.org/package=flextable

Gohel, D. (2021b). *officer: Manipulation of Microsoft Word and PowerPoint Documents*. R package version 0.3.19. https://CRAN.R-project.org/package=officer

Gremmen, M. C., Dijkstra, J. K., Steglich, C., & Veenstra, R. (2017). First selection, then influence: Developmental differences in friendship dynamics regarding academic achievement. *Developmental Psychology*, *53*(7), 1356–1370. https://doi.org/10.1037/dev0000314

Händel, M., Duan, X., & Vialle, W. (2018). Popular and smart? A cross-cultural study of students' perspectives on their peers. *Psychological Test and Assessment Modeling*, *60*(4), 429–449.

Händel, M., Vialle, W., & Ziegler, A. (2013). Student perceptions of high-achieving classmates. *High Ability Studies*, *24*(2), 99–114. https://doi.org/10.1080/13598139.2013.843139

Hannover, B., & Kessels, U. (2004). Self-to-prototype matching as a strategy for making academic choices. Why high school students do not like math and science. *Learning and Instruction*, *14*(1), 51–67. https://doi.org/10.1016/j.learninstruc.2003.10.002

Hannover, B., & Zander, L. (2020). How Personal and Social Selves Influence the Development of Children and Adolescents at School. *Zeitschrift Für Pädagogische Psychologie*, *34*(2), 65–85. https://doi.org/10.1024/1010-0652/a000261

Harris, K., & Vazire, S. (2016). On friendship development and the Big Five personality traits. *Social and Personality Psychology Compass*, *10*(11), 647–667. https://doi.org/10.1111/spc3.12287

Henry, L., & Wickham, H. (2020). *purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr

Hothorn, T., Zeileis, A., Farebrother, R. W., & Cummins, C. (2020). *lmtest: Testing Linear Regression Models*. https://CRAN.R-project.org/package=lmtest

Huber, C. (2011). Lehrerfeedback und soziale Integraton. Wie soziale Referenzierungsprozesse die soziale Integration in der Schule beeinflussen können [Teacher´s feedback and social integration: Is there a link between social referencing theory and social integration in school?]. *Empirische Sonderpädagogik*, *3*(1), 20–36.

Jansen, M [Malte], Schroeders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences. *Learning and Individual Differences*, *30*, 11–21. https://doi.org/10.1016/j.lindif.2013.12.003

Jetten, J., & Hornsey, M. J. (2014). Deviance and dissent in groups. *Annual Review of Psychology*, *65*, 461–485. https://doi.org/10.1146/annurev-psych-010213-115151

Jones, S., & Myhill, D. (2004). 'Troublesome boys' and 'compliant girls': gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education*, *25*(5), 547–561. https://doi.org/10.1080/0142569042000252044

Kessels, U. (2005). Fitting into the stereotype: How gender-stereotyped perceptions of prototypic peers relate to liking for school subjects. *European Journal of Psychology of Education*, *20*(3), 309–323. http://www.jstor.org/stable/23421531

Kessels, U., Heyder, A., Latsch, M., & Hannover, B. (2014). How gender differences in academic engagement relate to students' gender identity. *Educational Research*, *56*(2), 220–229. https://doi.org/10.1080/00131881.2014.898916

Kihlstrom, J. F., & Cantor, N. (2011). Social Intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (pp. 564–581). Cambridge University Press.

Kleiser, M., & Mayeux, L. (2021). Popularity and Gender Prototypicality: An Experimental Approach. *Journal of Youth and Adolescence*, *50*(1), 144–158. https://doi.org/10.1007/s10964-020-01344-5

Koenker, R. (2021). *SparseM: Sparse Linear Algebra*. http://www.econ.uiuc.edu/ roger/research/sparse/sparse.html

Köller, O., & Baumert, J. (2017). Hochleistende Schülerinnen und Schüler im mehr-und zweigliedrigen System [High-achieving students in the multipartite and two-tier systems.]. In M. Neumann, M. Becker, J. Baumert, K. Maaz, & O. Köller (Eds.), *Zweigliedrigkeit im deutschen Schulsystem: Potenziale und Herausforderungen in Berlin* (1st ed., p. 227). Waxmann Verlag GmbH.

Koster, M., Nakken, H., Pijl, S. J., & van Houten, E. (2009). Being part of the peer group: a literature study focusing on the social dimension of inclusion in education. *International Journal of Inclusive Education*, *13*(2), 117–140. https://doi.org/10.1080/13603110701284680

Krawinkel, S., Südkamp, A., Lange, S., & Tröster, H. (2017). Soziale Partizipation in inklusiven Grundschulklassen: Bedeutung von Klassen- und Lehrkraftmerkmalen [Social participation in inclusive classrooms: Relevance of classroom andteacher characteristics]. *Empirische Sonderpädagogik*, *3*, 277–295.

Kruse, H., & Kroneberg, C. (2020). *Contextualizing oppositional cultures: A multilevel network analysis of status orders in schools* (ECONtribute Discussion Papers Series No. 44). University of Bonn and University of Cologne, Germany. https://ideas.repec.org/p/ajk/ajkdps/044.html

Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1997). Classroom peer acceptance, friendship, and victimization: Distinct relational systems that contribute uniquely to children's school adjustment? *Child Development*, *68*(6), 1181–1197. https://doi.org/10.1111/j.1467-8624.1997.tb01993.x

Legewie, J., & DiPrete, T. A. (2012). School Context and the Gender Gap in Educational Achievement. *American Sociological Review*, *77*(3), 463–485. https://doi.org/10.1177/0003122412440802

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, *9*(1), 1–19.

Lumley, T. (2010). Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R. John Wiley and Sons.

Lumley, T. (2021). *survey: Analysis of Complex Survey Samples*. http://r-survey.r-forge.r-project.org/survey/

Lummis, M., & Stevenson, H. W. (1990). Gender differences in beliefs and achievement: A cross-cultural study. *Developmental Psychology*, *26*(2), 254–263. https://doi.org/10.1037/0012-1649.26.2.254

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: how explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*(2), 99–106. https://doi.org/10.1111/j.1750-8606.2008.00048.x

Maccoby, E. E. (1990). Gender and relationships: A developmental account. *American Psychologist*, *45*(4), 513–520. https://doi.org/10.1037/0003-066X.45.4.513

Masters, S. L., Hixson, K., & Hayes, A. R. (2021). Perceptions of Gender Norm Violations Among Middle School Students: An Experimental Study of the Effects of Violation Type on Exclusion Expectations. *The Journal of Early Adolescence*, *41*(4), 527–549. https://doi.org/10.1177/0272431620931193

Masur, P. K., & Scharkow, M. (2019). specr: Statistical functions for conducting specification curve analyses (Version 0.2.1). https://github.com/masurp/specr

Mayeux, L., & Kleiser, M. (2020). A Gender Prototypicality Theory of Adolescent Peer Popularity. *Adolescent Research Review*, *5*(3), 295–306. https://doi.org/10.1007/s40894-019-00123-z

Meijs, N., Cillessen, A. H. N., Scholte, R. H. J., Segers, E., & Spijkerman, R. (2010). Social intelligence and academic achievement as predictors of adolescent popularity. *Journal of Youth and Adolescence*, *39*(1), 62–72. https://doi.org/10.1007/s10964-008-9373-9

Müller, K., & Wickham, H. (2021). *tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble

Muntoni, F., & Retelsdorf, J. (2019). At their children‘s expense: How parents‘ gender stereotypes affect their children‘s reading outcomes. *Learning and Instruction*, *60*, 95–103. https://doi.org/10.1016/j.learninstruc.2018.12.002

OECD (2019). *PISA 2018 Results (Volume II).* OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en

Oehlschlägel, J., & Ripley, B. (2020). *bit: Classes and Methods for Fast Memory-Efficient Boolean Selections*. https://github.com/truecluster/bit

Olweus, D. (1991). Chapter 3 Victimization Among School Children. In R. Baenninger (Ed.), *Advances in Psychology: Vol. 76. Targets of Violence and Aggression* (1st ed., Vol. 76, pp. 45–102). Elsevier textbooks. https://doi.org/10.1016/S0166-4115(08)61056-0

Osterman, K. F. (2000). Students‘ Need for Belonging in the School Community. *Review of Educational Research*, *70*(3), 323–367. https://www.jstor.org/stable/1170786

Palacios, D., Dijkstra, J. K., Villalobos, C., Treviño, E., Berger, C., Huisman, M., & Veenstra, R. (2019). Classroom ability composition and the role of academic performance and school misconduct in the formation of academic and friendship networks. *Journal of School Psychology*, *74*, 58–73. https://doi.org/10.1016/j.jsp.2019.05.006

Pebesma, E., & Bivand, R. (2021). *sp: Classes and Methods for Spatial Data*. https://CRAN.R-project.org/package=sp

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13. https://CRAN.R-project.org/doc/Rnews/

Pelkner, A.-K., & Boehnke, K [K.] (2003). Streber als Leistungsverweigerer? Projektidee und erstes Datenmaterial einer Studie zu mathematischen Schulleistungen [Nerds as refusers of performance?]. *Zeitschrift Für Erziehungswissenschaft*, *6*(1), 106–125.

Pinheiro, J., Bates, D., & R-core. (2021). *nlme: Linear and Nonlinear Mixed Effects Models*. https://svn.r-project.org/R-packages/trunk/nlme/

Putarek, V., & Keresteš, G. (2016). Self-perceived popularity in early adolescence. *Journal of Social and Personal Relationships*, *33*(2), 257–274. https://doi.org/10.1177/0265407515574465

R Core Team. (2021). *R: A language and environment for statistical computing.* (Version 4.1.1.) [Computer software]. R Foundation for Statistical Computing. Vienna. https://www.R-project.org

Rentzsch, K., Schröder–Abé, M., & Schütz, A. (2013). Being Called A 'Streber': The Roles of Personality and Competition in the Labelling of Academically Oriented Students. *European Journal of Personality*, *27*(5), 411–423. https://doi.org/10.1002/per.1884

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. https://personality-project.org/r/psych/ https://personality-project.org/r/psych-manual.pdf

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, *29*(2), 173–191. https://doi.org/10.1016/j.socnet.2006.08.002

Rose, A. J., Glick, G. C., & Smith, R. L. (2011). Popularity and gender: The two cultures of boys and girls. In A. H. N. Cillessen, D. Schwartz, & L. Mayeux (Eds.), *Popularity in the Peer System* (pp. 103–122). Guilford Press.

Rose, A. J., & Rudolph, K. D. (2006). A review of sex differences in peer relationship processes: Potential trade-offs for the emotional and behavioral development of girls and boys. *Psychological Bulletin*, *132*(1), 98–131. https://doi.org/10.1037/0033-2909.132.1.98

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2021). *lavaan: Latent Variable Analysis*. https://lavaan.ugent.be

Rubin, K. H., Bukowski, W. M., & Parker, J. G. (2006). Peer Interactions, Relationships, and Groups. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (6th ed.). Wiley. https://doi.org/10.1002/9780470147658.chpsy0310

Rueger, S. Y., Malecki, C. K., & Demaray, M. K. (2008). Gender differences in the relationship between perceived social support and student adjustment during early adolescence. *School Psychology Quarterly*, *23*(4), 496–514. https://doi.org/10.1037/1045-3830.23.4.496

Sarkar, D. (2021). *lattice: Trellis graphics for R*. http://lattice.r-forge.r-project.org/

Selfhout, M., Burk, W., Branje, S., Denissen, J., van Aken, M., & Meeus, W. (2010). Emerging late adolescent friendship networks and Big Five personality traits: A social network approach. *Journal of Personality*, *78*(2), 509–538. https://doi.org/10.1111/j.1467-6494.2010.00625.x

Shin, H., & Ryan, A. M. (2014). Early adolescent friendships and academic adjustment: Examining selection and influence processes with longitudinal social network analysis. *Developmental Psychology*, *50*(11), 2462–2472. https://doi.org/10.1037/a0037922

Signorell, A. (2021). *DescTools: Tools for Descriptive Statistics*. https://CRAN.R-project.org/package=DescTools

Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science*, *1*(4), 538–555. https://doi.org/10.1177/2515245918805755

Simonsohn, U., & Gruson, H. (2021). *groundhog: The Simplest Solution to Version-Control for CRAN Packages*. https://CRAN.R-project.org/package=groundhog

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Smith, T. E., & Leaper, C. (2006). Self-Perceived Gender Typicality and the Peer Context During Adolescence. *Journal of Research on Adolescence*, *16*(1), 91–104. https://doi.org/10.1111/j.1532-7795.2006.00123.x

Stadtfeld, C., Vörös, A., Elmer, T., Boda, Z., & Raabe, I. J. (2019). Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(3), 792–797. https://doi.org/10.1073/pnas.1811388115

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (Eds.). (2019a). IQB Trends in Student Achievement 2018. The second assessment of mathematics and science proficiencies at the end of ninth grade: Summary. Waxmann.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (Eds.). (2019b). IQB-Bildungstrend 2018: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich. Waxmann.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Therneau, T. M. (2021). *survival: Survival Analysis*. https://github.com/therneau/survival

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://www.jstatsoft.org/v45/i03/

van Buuren, S., & Groothuis-Oudshoorn, K. (2021). *mice: Multivariate Imputation by Chained Equations*. https://CRAN.R-project.org/package=mice

van Houtte, M. (2006). School type and academic culture: evidence for the differentiation–polarization theory. *Journal of Curriculum Studies*, *38*(3), 273–292. https://doi.org/10.1080/00220270500363661

Vannatta, K., Gartstein, M. A., Zeller, M., & Noll, R. B. (2009). Peer acceptance and social behavior during childhood and adolescence: How important are appearance, athleticism, and academic competence? *International Journal of Behavioral Development*, *33*(4), 303–311. https://doi.org/10.1177/0165025408101275

Visser, I. (1996). The prototypicality of gender: Contemporary notions of masculine and feminine. *Women's Studies International Forum*, *19*(6), 589–600.

Vörös, A., Block, P., & Boda, Z. (2019). Limits to inferring status from friendship relations. *Social Networks*, *59*, 77–97. https://doi.org/10.1016/j.socnet.2019.05.007

Wang, M.-T., & Degol, J. (2013). Motivational Pathways to STEM Career Choices: Using Expectancy-Value Perspective to Understand Individual and Gender Differences in STEM Fields. *Developmental Review : DR*, *33*(4). https://doi.org/10.1016/j.dr.2013.08.001

Weirich, S., Hecht, M., & Becker, B. (2021). *eatRep: Educational Assessment Tools for Replication Methods*. https://github.com/weirichs/eatRep

Wentzel, K. R. (1991). Relations between Social Competence and Academic Achievement in Early Adolescence. *Child Development*, *62*(5), 1066–1078.

Wentzel, K. R., & Asher, S. R. (1995). The academic lives of neglected, rejected, popular, and controversial children. *Child Development*, *66*(3), 754–763. https://doi.org/10.1111/j.1467-8624.1995.tb00903.x

Wentzel, K. R., Barry, C. M., & Caldwell, K. A. (2004). Friendships in middle school: Influences on motivation and school adjustment. *Journal of Educational Psychology*, *96*(2), 195–203. https://doi.org/10.1037/0022-0663.96.2.195

Wentzel, K. R., Jablansky, S., & Scalise, N. R. (2021). Peer social acceptance and academic achievement: A meta-analytic study. *Journal of Educational Psychology*, *113*(1), 157–180. https://doi.org/10.1037/edu0000468

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Wilke, C. O. (2020). cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2. https://wilkelab.org/cowplot/

Wolter, I., & Seidel, T. (2017). Kompetent und beliebt? Der Zusammenhang von Kompetenzen und Selbstkonzepten in Mathematik und Lesen mit der wahrgenommenen Beliebtheit bei Peers [Competent and popular? The relationship of competencies and self-concepts in mathematics and reading to perceived peer popularity]. *Zeitschrift Für Erziehungswissenschaft*, *20*(3), 387–404. https://doi.org/10.1007/s11618-017-0772-0

Wood, S. (2021). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. https://CRAN.R-project.org/package=mgcv

Workman, J., & Heyder, A. (2020). Gender achievement gaps: the role of social costs to trying hard in high school. *Social Psychology of Education*, *23*(6), 1407–1427. https://doi.org/10.1007/s11218-020-09588-6

Youniss, J., & Haynie, D. L. (1992). Friendship in adolescence. *Developmental and Behavioral Pediatrics*, *13*(1), 59–66.

Zeileis, A., & Croissant, Y. (2020). *Formula: Extended Model Formulas*. https://CRAN.R-project.org/package=Formula

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, *14*(6), 1–27. https://doi.org/10.18637/jss.v014.i06

Zeileis, A., Grothendieck, G., & Ryan, J. A. (2021). *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*. https://zoo.R-Forge.R-project.org/

Zeileis, A., & Lumley, T. (2021). *sandwich: Robust Covariance Matrix Estimators*. https://sandwich.R-Forge.R-project.org/

Appendix III-A: Gender comparisons

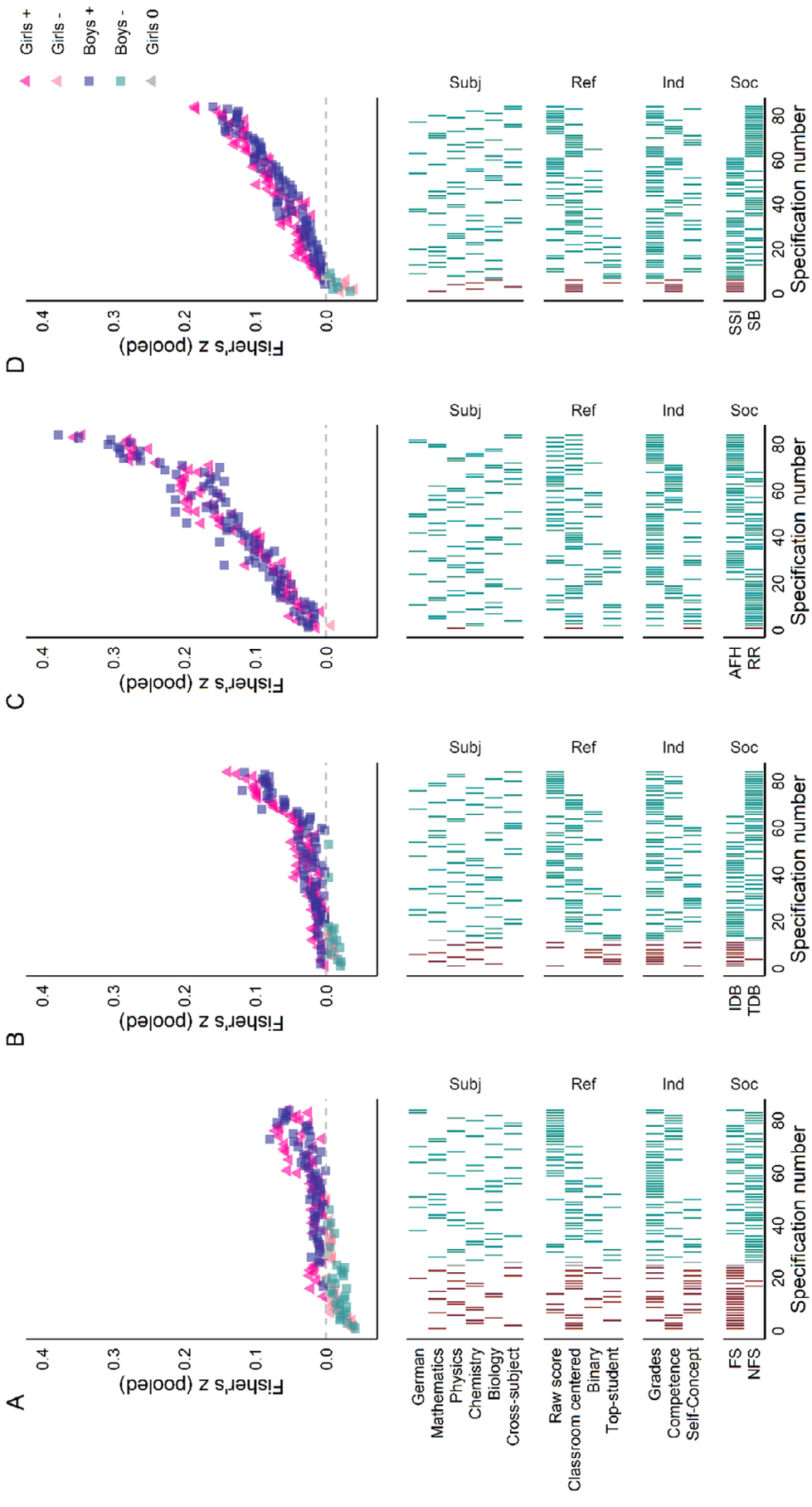


Figure III-A1. Specification Curves of the Correlation Between Achievement and Integration by Gender. For each facet of social integration, a specification curve is depicted which shows the correlation between achievement and social integration. The top panel shows the distribution of Fisher's Z values across the combinations of achievement and social integration indicator (i.e., the specifications). The middle panel indicates the specification belonging to the respective value in the top panel. Achievement indicators are characterized by a unique configuration of the subject of achievement ("Subj"), the reference ("Ref") and the indicator ("Ind") chosen. Social integration indicators ("Soc") are explained below. The bottom panel depicts the marginal distribution of Fisher's Z by each specification factor. Panel A: Friendship. NFS = No reciprocal friendship nominations, FS = Reciprocal friendship nominations. Panel B: Contacts. TDB = Spending time during breaks together in-degrees, IDB = Social isolation during breaks. Panel C: Peer acceptance. RR = Rejection nominations (recoded), AFH = Being asked for help nominations. Panel D: Self perceived acceptance. SSI = Self-perceived social integration, SB = School belonging.

Table III-A1

Two-Lines Test for Non-Linear Relationship Between Achievement and Different Dimensions of Social Integration Performed for Boys and Girls Separately.

Dimensions	Boys							Girls						
	Slope 1			Slope 2			Cut-point	Slope 1			Slope 2			Cut-point
	b_1	t	p	b_2	t	p	x_{orig}	b_1	t	p	b_2	t	p	x_{orig}
Reciprocal friendship nominations	0.02	6.77	***	-0.01	-1.55		2.80	0.02	9.49	***	-0.01	-1.39		2.40
Acceptance nominations	0.02	13.90	***	0.00	-0.12		1.61	0.03	11.71	***	0.01	0.38		1.83
Being asked for help nominations	0.04	18.13	***	0.10	24.85	***	2.80	0.06	44.42	***	0.08	2.04	*	1.40
Rejection nominations (recoded)	0.04	24.62	***	0.01	0.32		1.45	0.04	24.21	***	-0.03	-1.06		1.40
Contact out-degrees	0.01	7.20	***	-0.03	-1.07		1.77	0.02	10.78	***	0.01	0.33		1.72
Self-perceived social integration	0.07	9.79	***	-0.01	-0.13		1.81	0.08	13.89	***	-0.03	-0.39		1.59
School belonging	0.11	19.03	***	-0.05	-0.72		1.72	0.14	21.68	***	-0.04	-0.81		1.68

Note. * p < .05, ** p < .01, *** p < .001.

Appendix III-B: Regression Tables

Table III-B1

Reciprocal Friendship Nominations Regressed on Average Grade and Gender

Variable	(1)			(2)			(3)			(4)		(5)	
	B	*p*	β	*B*	*p*	β	*B*	*p*	β	*B*	*p*	*B*	*p*
Intercept	0.268	***		0.267	***		0.263	***		-1.046	***	-1.058	***
Average Grade (linear term)	0.005	***	0.032	-0.009	**	-0.058	-0.013	**	-0.083	0.028	***	0.016	
Average Grade (quadratic term)				-0.008	***	-0.051	-0.008	***	-0.051				
Gender (girls)	0.001		0.006	0.001		0.006	0.007		0.045	0.003		0.025	
Average Grade*Gender							0.005		0.032			0.024	
Average Grade (quadratic term)*Gender							-0.001		-0.006				
Covariates													
Network density	0.836	***	5.356	0.836	***	5.356	0.835	***	5.349	4.303	***	4.301	***
Class size	-0.001		-0.006	-0.001		-0.006	-0.001		-0.006	-0.001		-0.001	
School type (highest track)	0.004		0.026	0.004		0.026	0.004		0.026	0.026	*	0.026	*
School type (lowest track)	-0.014	**	-0.090	-0.014	**	-0.090	-0.014	**	-0.090	-0.092	**	-0.092	**
HISEI	0.000	**	0.000	0.000	*	0.000	0.000	*	0.000	-0.001	**	-0.001	**
Language	0.002		0.013	0.001		0.006	0.001		0.006	0.013		0.012	
R-squared	0.292			0.295			0.295			0.304		0.304	

Note. Models 1 - 3 are linear probability models, Models 4 and 5 are quasibinomial with logit link. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992).
* p < .05, ** p < .01, *** p < .001

Table III-B2

Reciprocal friendship nominations Regressed on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)			(3)			(4)		
	B	*p*	*β*	*B*	*p*	*β*	*B*	*p*	*OR*	*B*	*p*	*OR*
Intercept	0.253	***		0.248	***		-1.124	***		-1.158	***	
Top grades in neither domain	0.008		0.051	0.014		0.090	0.043		1.044	0.080		1.083
Top grades in German/Biology	0.010		0.064	0.021		0.135	0.053		1.054	0.117		1.124
Top grades in Math/Physics	0.005		0.032	0.005		0.032	0.028		1.028	0.033		1.034
Gender (girls)	0.002		0.013	0.011		0.070	0.009		1.009	0.064		1.066
Top grades in neither domain*Gender				-0.010		-0.064				-0.059		0.943
Top grades in German/Biology*Gender				-0.017		-0.109				-0.098		0.907
Top grades in Math/Physics*Gender				0.002		0.013				0.004		1.004
Covariates												
Network density	0.837	***	5.362	0.837	***	5.362	4.311	***	74.515	4.310	***	74.440
Class size	0.000		0.000	0.000		0.000	-0.001		0.999	-0.001		0.999
School type (highest track)	0.006	*	0.038	0.006	*	0.038	0.033	*	1.034	0.033	*	1.034
School type (lowest track)	-0.014	**	-0.090	-0.014	**	-0.090	-0.091	**	0.913	-0.090	**	0.914
HISEI	0.000	*	0.000	0.000	*	0.000	-0.001	*	0.999	-0.001	*	0.999
Language	0.003		0.019	0.003		0.019	0.019		1.019	0.020		1.020
R-squared	0.292			0.292			0.304			0.304		

Note. Models 1 and 2 are linear probability models, Models 3 and 4 are quasibinomial with logit link. β has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor. Average grade has been centered at 4 (equivalent to German Grade 2 out of 6).

* p < .05, ** p < .01, *** p < .001

Table III-B3

Contact During Breaks (In-Degrees) Regressed on Average Grade and Gender

Variable	(1)			(2)			(3)			(4)		(5)	
	B	*p*	β	*B*	*p*	β	*B*	*p*	β	*B*	*p*	*B*	*p*
Intercept	0.235	***		0.235	***		0.236	***		-1.205	***	-1.206	***
Average Grade (linear term)	0.013	***	0.109	0.006	*	0.050	0.010	**	0.084	0.079	***	0.079	***
Average Grade (quadratic term)				-0.004	***	-0.033	-0.002		-0.017				
Gender (girls)	-0.023	***	-0.192	-0.023	***	-0.192	-0.024	***	-0.201	-0.141	***	-0.141	***
Average Grade*Gender							-0.008		-0.067			0.000	
Average Grade (quadratic term)*Gender							-0.004		-0.033				
Covariates													
Network density	0.886	***	7.406	0.886	***	7.406	0.886	***	7.406	4.970	***	4.970	***
Class size	-0.001	*	-0.008	-0.001	*	-0.008	-0.001	*	-0.008	-0.002		-0.002	
School type (highest track)	-0.012	***	-0.100	-0.012	***	-0.100	-0.012	***	-0.100	-0.062	***	-0.062	***
School type (lowest track)	0.009	*	0.075	0.010	*	0.084	0.010	*	0.084	0.057		0.057	
HISEI	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	
Language	0.006	*	0.050	0.005	*	0.042	0.005	*	0.042	0.036	*	0.036	*
R-squared	0.234			0.236	***		0.236			0.236	***	0.236	***

Note. Models 1 - 3 are linear probability models, Models 4 and 5 are quasibinomial with logit link. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992).

* p < .05, ** p < .01, *** p < .001

Table III-B4

Contact during breaks (In-Degrees) Regressed on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)			(3)			(4)		
	B	*p*	*β*	*B*	*p*	*β*	*B*	*p*	*OR*	*B*	*p*	*OR*
Intercept	0.229	***		0.233	***		-1.248	***		-1.231	***	
Top grades in neither domain	-0.011	*	-0.092	-0.015	*	-0.125	-0.062	**	0.940	-0.081	*	0.922
Top grades in German/Biology	-0.002		-0.017	0.001		0.008	-0.014		0.986	0.007		1.007
Top grades in Math/Physics	-0.003		-0.025	-0.007		-0.059	-0.014		0.986	-0.030		0.970
Gender (girls)	-0.021	***	-0.176	-0.027	**	-0.226	-0.127	***	0.881	-0.155	***	0.856
Top grades in neither domain*Gender				0.007		0.059				0.032		1.033
Top grades in German/Biology*Gender				-0.005		-0.042				-0.030		0.970
Top grades in Math/Physics*Gender				0.006		0.050				0.027		1.027
Covariates												
Network density	0.895	***	7.482	0.895	***	7.482	5.019	***	151.260	5.021	***	151.563
Class size	0.000		0.000	0.000		0.000	-0.002		0.998	-0.002		0.998
School type (highest track)	-0.010	***	-0.084	-0.010	***	-0.084	-0.051	***	0.950	-0.051	***	0.950
School type (lowest track)	0.009	*	0.075	0.010	*	0.084	0.059		1.061	0.059		1.061
HISEI	0.000		0.000	0.000		0.000	0.000		1.000	0.000		1.000
Language	0.008	***	0.067	0.008	***	0.067	0.050	***	1.051	0.050	***	1.051
R-squared	0.229			0.229			0.231			0.231		

Note. Models 1 and 2 are linear probability models, Models 4 and 5 are quasibinomial with logit link. 'β' has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor. Average grade has been centered at 4 (equivalent to German Grade 2 out of 6).

* p < .05, ** p < .01, *** p < .001

Table III-B5

Being Asked for Help In-degrees Regressed on Average Grade and Gender

Variable	(1)			(2)			(3)			(4)		(5)	
	B	*p*	β	*B*	*p*	β	*B*	*p*	β	*B*	*p*	*B*	*p*
Intercept	0.249	***		0.253	***		0.257	***		-1.124	***	-1.092	***
Average Grade (linear term)	0.056	***	0.446	0.082	***	0.653	0.093	***	0.740	0.359	***	0.394	***
Average Grade (quadratic term)				0.014	***	0.111	0.018	***	0.143				
Gender (girls)	0.032	***	0.255	0.032	***	0.255	0.025	***	0.199	0.203	***	0.148	***
Average Grade*Gender							-0.019	***	-0.151			-0.066	***
Average Grade (quadratic term)*Gender							-0.007	**	-0.056				
Covariates													
Network density	0.818	***	6.512	0.822	***	6.544	0.823	***	6.552	4.973	***	4.978	***
Class size	-0.001	*	-0.008	-0.001		-0.008	-0.001	*	-0.008	-0.003		-0.003	
School type (highest track)	-0.028	***	-0.223	-0.028	***	-0.223	-0.028	***	-0.223	-0.170	***	-0.170	***
School type (lowest track)	0.008	*	0.064	0.007		0.056	0.007		0.056	0.050	*	0.050	*
HISEI	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	
Language	-0.003		-0.024	-0.003		-0.024	-0.003		-0.024	-0.023		-0.022	
R-squared	0.286			0.291	***		0.291			0.289	***	0.288	***

Note. Models 1 - 3 are linear probability models, Models 4 and 5 are quasibinomial with logit link. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992).

* p < .05, ** p < .01, *** p < .001

Table III-**B6**

Being Asked for Help Regressed on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)			(3)			(4)		
	B	p	β	B	p	β	B	p	OR	B	p	OR
Intercept	0.294	***		0.307	***		-0.935	***		-0.824	***	
Top grades in neither domain	-0.122	***	-0.971	-0.137	***	-1.091	-0.664	***	0.515	-0.791	***	0.453
Top grades in German/Biology	-0.061	***	-0.486	-0.072	***	-0.573	-0.306	***	0.736	-0.378	***	0.685
Top grades in Math/Physics	-0.054	***	-0.430	-0.054	***	-0.430	-0.260	***	0.771	-0.276	***	0.759
Gender (girls)	0.039	***	0.310	0.017		0.135	0.245	***	1.278	0.072		1.075
Top grades in neither domain*Gender				0.024	*	0.191				0.205	***	1.228
Top grades in German/Biology*Gender				0.018		0.143				0.117		1.124
Top grades in Math/Physics*Gender				-0.005		-0.040				-0.010		0.990
Covariates												
Network density	0.849	***	6.759	0.850	***	6.767	5.142	***	171.058	5.147	***	171.915
Class size	-0.001		-0.008	-0.001	*	-0.008	-0.002		0.998	-0.002		0.998
School type (highest track)	-0.023	***	-0.183	-0.023	***	-0.183	-0.136	***	0.873	-0.136	***	0.873
School type (lowest track)	0.009	*	0.072	0.009	*	0.072	0.059	*	1.061	0.060	*	1.062
HISEI	0.000	*	0.000	0.000	*	0.000	0.001	*	1.001	0.001	*	1.001
Language	0.004	*	0.032	0.004	*	0.032	0.028	*	1.028	0.028	*	1.028
R-squared	0.235			0.236			0.237			0.238		

Note. Models 1 and 2 are linear probability models, Models 3 and 4 are quasibinomial with logit link. 'β' has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor. Average grade has been centered at 4 (equivalent to German Grade 2 out of 6).

* p < .05, ** p < .01, *** p < .001

Table III-**B7**

Rejection In-Degrees Regressed on Average Grade and Gender

Variable	(1)			(2)			(3)			(4)		(5)	
	B	*p*	β	*B*	*p*	β	*B*	*p*	β	*B*	*p*	*B*	*p*
Intercept	0.813	***		0.811	***		0.814	***		1.529	***	1.530	***
Average Grade (linear term)	0.024	***	0.176	0.013	***	0.096	0.018	***	0.132	0.165	***	0.166	***
Average Grade (quadratic term)				-0.006	***	-0.044	-0.004	*	-0.029				
Gender (girls)	0.018	***	0.132	0.018	***	0.132	0.015	***	0.110	0.130	***	0.128	***
Average Grade*Gender							-0.009	*	-0.066			-0.002	
Average Grade (quadratic term)*Gender							-0.004		-0.029				
Covariates													
Network density	-0.960	***	-7.056	-0.959	***	-7.048	-0.960	***	-7.056	-6.045	***	-6.045	***
Class size	0.000		0.000	0.000	*	0.000	0.000	*	0.000	-0.004	*	-0.004	*
School type (highest track)	0.007	***	0.051	0.007	***	0.051	0.007	***	0.051	0.045	***	0.045	***
School type (lowest track)	-0.015	***	-0.110	-0.015	***	-0.110	-0.015	***	-0.110	-0.089	**	-0.089	**
HISEI	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	
Language	0.006	**	0.044	0.005	**	0.037	0.005	**	0.037	0.039	**	0.039	**
R-squared	0.258			0.259	***		0.259			0.287	***	0.288	***

Note. Models 1 - 3 are linear probability models, Models 4 and 5 are quasibinomial with logit link. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992).

* p < .05, ** p < .01, *** p < .001

Table III-**B8**

Rejection In-Degrees Regressed on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)			(3)			(4)		
	B	*p*	*β*	*B*	*p*	*β*	*B*	*p*	*OR*	*B*	*p*	*OR*
Intercept	0.803***			0.808***			1.475***			1.497***		
Top grades in neither domain	-0.023***		-0.169	-0.029***		-0.213	-0.173***		0.841	-0.198***		0.820
Top grades in German/Biology	-0.006		-0.044	-0.010		-0.073	-0.041		0.960	-0.077		0.926
Top grades in Math/Physics	-0.006		-0.044	-0.002		-0.015	-0.049		0.952	-0.015		0.985
Gender (girls)	0.022***		0.162	0.014		0.103	0.155***		1.168	0.119		1.126
Gender*Top grades in neither domain				0.010		0.073				0.044		1.045
Gender*Top grades in German/Biology				0.008		0.059				0.059		1.061
Gender*Top grades in Math/Physics				-0.010		-0.073				-0.082		0.921
Covariates												
Network density	-0.963***		-7.078	-0.963***		-7.078	-6.056***		0.002	-6.058***		0.002
Class size	0.000		0.000	0.000		0.000	-0.003*		0.997	-0.003*		0.997
School type (highest track)	0.010***		0.073	0.010***		0.073	0.071***		1.074	0.071***		1.074
School type (lowest track)	-0.015***		-0.110	-0.015***		-0.110	-0.087**		0.917	-0.087**		0.917
HISEI	0.000		0.000	0.000		0.000	0.000		1.000	0.000		1.000
Language	0.010***		0.073	0.010***		0.073	0.068***		1.070	0.068***		1.070
R-squared	0.243			0.244			0.271			0.271		

Note. Models 1 and 2 are linear probability models, Models 3 and 4 are quasibinomial with logit link. 'β' has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor. Average grade has been centered at 4 (equivalent to German Grade 2 out of 6).

* p < .05, ** p < .01, *** p < .001

Table III- **B9**

Regression of Self-perceived Social Integration on Achievement and Gender

Variable	(1)			(2)			(3)		
	B	*p*	β	*B*	*p*	β	*B*	*p*	β
Intercept	3.230	***		3.228	***		3.220	***	
Average Grade (linear term)	0.046	***	0.091	0.031	**	0.061	0.013		0.026
Average Grade (quadratic term)				-0.008		-0.016	-0.014		-0.028
Gender (girls)	-0.004		-0.008	-0.004		-0.008	0.009		0.018
Average Grade*Gender							0.031		0.061
Average Grade (quadratic term)*Gender							0.011		0.022
Covariates									
Class size	0.003	*	0.006	0.003	*	0.006	0.003	*	0.006
School type (highest track)	0.080	***	0.158	0.079	***	0.156	0.079	***	0.156
School type (lowest track)	-0.009		-0.018	-0.009		-0.018	-0.009		-0.018
HISEI	0.000		0.000	0.000	*	0.000	0.000	*	0.000
Language	0.053	***	0.105	0.053	***	0.105	0.053	***	0.105
R-squared	0.019			0.019			0.020		

Note. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992).

* p < .05, ** p < .01, *** p < .001

Table III-B10

Self-Perceived Social Integration Regressed on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)		
	B	*p*	*β*	*B*	*p*	*β*
Intercept	3.212***			3.195***		
Top grades in neither domain	-0.043*		-0.085	-0.026		-0.051
Top grades in German/Biology	0.004		0.008	0.024		0.047
Top grades in Math/Physics	-0.028		-0.055	-0.014		-0.028
Gender (girls)	0.003		0.006	0.029		0.057
Top grades in neither domain*Gender				-0.028		-0.055
Top grades in German/Biology*Gender				-0.032		-0.063
Top grades in Math/Physics*Gender				-0.022		-0.044
Covariates						
Class size	0.003*		0.006	0.003*		0.006
School type (highest track)	0.087***		0.172	0.087***		0.172
School type (lowest track)	-0.009		-0.018	-0.009		-0.018
HISEI	0.001**		0.002	0.001**		0.002
Language	0.061***		0.121	0.061***		0.121
R-squared	0.015			0.015		

Note.'β' has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor.

* p < .05, ** p < .01, *** p < .001

Table III-B11

Regression of School Belonging on Achievement and Gender

Variable	(1)			(2)			(3)		
	B	*p*	β	*B*	*p*	β	*B*	*p*	β
Intercept	3.224***			3.223***			3.211***		
Average Grade (linear term)	0.110***		0.212	0.098***		0.189	0.076***		0.146
Average Grade (quadratic term)				-0.006		-0.012	-0.013		-0.025
Gender (girls)	-0.016*		-0.031	-0.016*		-0.031	0.003		0.006
Average Grade*Gender							0.038		0.073
Average Grade (quadratic term)*Gender							0.012		0.023
Covariates									
Class size	0.004***		0.008	0.004***		0.008	0.004***		0.008
School type (highest track)	0.047***		0.090	0.047***		0.090	0.047***		0.090
School type (lowest track)	0.018		0.035	0.019		0.037	0.019		0.037
HISEI	0.000		0.000	0.000		0.000	0.000		0.000
Language	0.055***		0.106	0.055***		0.106	0.055***		0.106
R-squared	0.039			0.039			0.039		

Note. Average grade has been centered at 4 (equivalent to German grade 2 out of 6). The following variables have been centered at their sample mean prior to analysis: Network density, class size, HISEI (highest parental international socio-economic index of occupational status; Ganzeboom et al., 1992).

* p < .05, ** p < .01, *** p < .001

Table III-B12

Linear Regression School Belonging on Stereotyped Subject Combinations and Gender

Variable	(1)			(2)		
	B	*p*	*β*	*B*	*p*	*β*
Intercept	3.221	***		3.201	***	
Top grades in neither domain	-0.146	***	-0.281	-0.126	***	-0.242
Top grades in German/Biology	-0.047		-0.090	-0.014		-0.027
Top grades in Math/Physics	-0.045	*	-0.087	-0.024		-0.046
Gender (girls)	0.000		0.000	0.032		0.062
Top grades in neither domain*Gender				-0.032		-0.062
Top grades in German/Biology*Gender				-0.052		-0.100
Top grades in Math/Physics*Gender				-0.035		-0.067
Covariates						
Class size	0.004	***	0.008	0.004	***	0.008
School type (highest track)	0.062	***	0.119	0.062	***	0.119
School type (lowest track)	0.018		0.035	0.018		0.035
HISEI	0.001	**	0.002	0.001	**	0.002
Language	0.073	***	0.140	0.073	***	0.140
R-squared	0.020			0.020		

Note.'β' has been standardized by the standard deviation of the dependent variable and denotes the standardized change in y at one-unit increase in the predictor.

* p < .05, ** p < .01, *** p < .001

4.3. Zusammenfassung und Zwischenfazit aus Artikel III

Diese Studie hat gezeigt, dass der Zusammenhang zwischen sozialer Integration und unterschiedlichen Operationalisierungen von Leistung zwischen leicht negativen und moderat positiven Werten variiert. Dabei zeigen sich systematische Unterschiede sowohl zwischen den Facetten sozialer Integration (Freundschaft zeigt die geringsten, Akzeptanz die größten Korrelationswerte) als auch zwischen verschiedenen Dimensionen und Operationalisierungen von Leistung bzw. Leistungsstärke. So sind vorherige Befunde, die sich vor allem auf die Ablehnung von Leistungsstereotypen beziehen, am ehesten mit den empirischen Befunden in Klassen zu vereinbaren, in denen es lediglich eine/einen „High Achiever“ gibt, während Theorien, nach denen verbale Fähigkeiten bedeutsam für den sozialen Austausch und die soziale Integration sind, hier stärker bei Freundschaften zum Tragen kommen, da dort die Deutschnote besonders hervorsticht.

Für das Konstrukt der sozialen Integration lässt sich aus diesen Ergebnissen schlussfolgern, dass durchaus verschiedene Facetten unterschiedliche Zusammenhänge zu Leistungsvariablen zeigen, eine differenziertere Betrachtung also notwendig ist. In der bisherigen Forschung ist die Verbindung zwischen theoretischen Konzepten und Operationalisierung noch nicht zufriedenstellend erfolgt. Unterschiede in den Zusammenhängen zwischen den Indikatoren innerhalb einer Facette sprechen für die Weiterentwicklung und Schärfung der theoretischen Ausformulierung des Konzepts der sozialen Integration. Dabei ist in diesem Zusammenhang die Frage zu stellen, ob die Facetten, die einen unterschiedlichen theoretischen Bedeutungsgehalt haben, auch in der Empirie trennscharf operationalisiert werden können. Insbesondere für die beiden Facetten „Interaktion“ und „Akzeptanz“ stellt sich die Frage, wie eine valide Operationalisierung aussehen könnte. In der vorliegenden Arbeit haben wir Interaktionen über die Nominierungen, mit wem Schülerinnen und Schüler die Pausen verbringen operationalisiert, während Akzeptanz durch die Items zu Hilfeersuchen und unerwünschten Sitznachbarn operationalisiert. Diese Zuordnung könnte jedoch durchaus in Frage gestellt werden, da Akzeptanz gemeinhin eine Bedingung für Interaktionen ist, und umgekehrt das Eintreten in Interaktionen ein Indikator für Akzeptanz sein könnte.

Auch für die Forschung zum Thema Leistungs- und Geschlechterstereotype haben die Studienergebnisse Implikationen. Diese Untersuchung bietet nur begrenzte Evidenz für systematische Einflüsse des Geschlechts auf die leistungsabhängige soziale Integration. Für die Frage, wer durch Mitschülerinnen und Mitschüler um Hilfe gebeten wird, zeigte sich, dass Mädchen generell häufiger um Hilfe gebeten werden und auch, dass Jungen deutlich höhere Leistungen zeigen müssen, um ähnlich häufig um Hilfe gebeten zu werden. Hier könnte eventuell ein Zusammenhang zwischen Leistungsstärke und tatsächlichem oder wahrgenommenen

prosozialem Verhalten eine Rolle spielen, der möglicherweise bei Jungen stärker ausgeprägt ist. Diese Spekulation müsste jedoch durch spezifischere Forschung erhärtet werden. Von diesem Interaktionseffekt abgesehen zeigten sich jedoch keine weiteren konsistenten Geschlechterunterschiede bei der sozialen Integration leistungsstarker Schülerinnen und Schüler. Dies gilt sowohl für die mittlere Fachleistung als auch für geschlechtskonnotierte Leistungsprofile.

Die Ergebnisse ermutigen zunächst, da wir im Großen und Ganzen keine negativen Auswirkungen von Leistungsstärke auf die tatsächliche soziale Integration von Schülerinnen und Schülern finden. Im Gegenteil sind Schülerinnen und Schüler, die höhere Leistungen zeigen, im Schnitt auch besser sozial integriert. Folgende Einschränkungen können aber gemacht werden: *Erstens* beruhten unsere weiterführenden Analysen (über die Berechnung der Korrelation hinaus) auf Notenschnitten und nahmen daher eine übergreifende Perspektive ein. Der generelle Zusammenhang zwischen Leistung und sozialer Integration ist positiv, aber in den einzelnen Klassen könnte dies auch anders aussehen – gerade, wenn man sich in der Multiversumsanalyse anschaut, dass Bedingungen, in denen nur die Klassen gewählt wurden, in denen die Noten tatsächlich zwischen den Schülern und Schülerinnen diskriminieren, bzw. ein solches Leistungsgefälle vorherrscht, dass lediglich ein Schüler oder eine Schülerin an der Leistungsspitze steht. Für diese Substichprobe zeigten sich nämlich deutlich geringere und zum Teil negative Korrelationen mit Facetten sozialer Integration. Allerdings zeigen Studien, die auf Klassenebene sogenannte *oppositional cultures*, also Klassennormen, in denen Leistung abgewertet wird, untersuchen, dass diese in deutschen Schulklassen sehr selten sind (Kruse & Kroneberg, 2020; Lorenz, Boda & Salikutluk, 2021).

Zweitens könnte man weiterhin argumentieren, dass die Verwendung von Noten als Leistungsindikator die Stereotype Bewertung von Schülerinnen und Schülern, die in einem Fach leistungsstark sind, welches als inkongruent mit deren Geschlecht wahrgenommen wird, nicht das ganze Bild zeigt, da gerade in den Klassen, in denen dieses Urteil verbreitet wäre, Schülerinnen und Schüler sich nicht trauen, die Leistungen zu zeigen, zu denen sie fähig wären. Diese Annahme ging Boehnke (2008) nach. Indem er untersuchte, wie die Angst vor sozialer Ausgrenzung den Zusammenhang zwischen Noten und Kompetenzen im Fach Mathematik moderierte, zeigte er, dass besonders Mädchen anfällig für Gruppenzwang waren, das heißt, bei leistungsstarken Mädchen mit hoher Angst vor sozialer Ausgrenzung gab es eine ausgeprägte negative Korrelation zwischen Noten und Testleistungen. Einschränkend muss allerdings gesagt werden, dass die Gruppe derjenigen, die eine hohe Angst bei gleichzeitig hoher Leistung aufwiesen lediglich zwei

Prozent der Stichprobe ausmachten ($N = 35$) und nochmal kleiner wurde, als die Ergebnisse zwischen Jungen und Mädchen verglichen wurden.

Aus diesem Beitrag lassen sich weitere Fragestellungen für die Forschung ableiten. Zunächst wäre eine weitere Untersuchung und Differenzierung der Facetten von sozialer Integration von Interesse. Welche Facetten sind besonders bedeutsam für unterschiedliche Outcomes, wie Wohlbefinden, Verhalten und die Leistungsentwicklung von Schülerinnen und Schülern? Welche Bedingungen führen zu einer hohen Integration in unterschiedlichen Facetten? Wie verhalten sich die verschiedenen Facetten zueinander? Möglicherweise könnten auf Basis eines genaueren Verständnisses von sozialer Integration spezifischere Interventionen entwickelt werden, die die sozialen Beziehungen in Klassen durch verschiedene Ansatzpunkte fördern ließen. (Zander et al., 2017).

Literaturverzeichnis

Abrams, D. & Killen, M. (2014). Social Exclusion of Children: Developmental Origins of Prejudice. *Journal of Social Issues*, *70*(1), 1–11. https://doi.org/10.1111/josi.12043

Bergold, S., Kasper, D., Wendt, H. & Steinmayr, R. (2020). Being bullied at school: the case of high-achieving boys. *Social Psychology of Education*, *23*(2), 315–338. https://doi.org/10.1007/s11218-019-09539-w

Boehnke, K. (2008). Peer pressure: a cause of scholastic underachievement? A cross-cultural study of mathematical achievement among German, Canadian, and Israeli middle school students. *Social Psychology of Education*, *11*(2), 149–160. https://doi.org/10.1007/s11218-007-9041-z

Bossaert, G., Colpin, H., Pijl, S. J. & Petry, K. (2013). Truly included? A literature study focusing on the social dimension of inclusion in education. *International Journal of Inclusive Education*, *17*(1), 60–79. https://doi.org/10.1080/13603116.2011.580464

Brown, B. B. (2004). Adolescents' Relationships with Peers. In R. M. Lerner & L. D. Steinberg (Hrsg.), *Handbook of Adolescent Psychology* (S. 363–394). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Brown, B. B. & Larson, J. (2009). Peer relationships in adolescence. In R. M. Lerner & L. D. Steinberg (Eds.), *Handbook of adolescent psychology. Contextual influences on adolescent development* (vol. 2, S. 74–103). Hoboken, N. J.: John Wiley & Sons.

Cillessen, A. H. N., Schwartz, D. & Mayeux, L. (Hrsg.). (2011). *Popularity in the Peer System*: Guilford Press.

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Gest, S. D., Graham-Bermann, S. A. & Hartup, W. W. (2001). Peer Experience: Common and Unique Features of Number of Friendships, Social Network Centrality, and Sociometric Status. *Social Development*, *10*(1), 23–40. https://doi.org/10.1111/1467-9507.00146

Gifford-Smith, M. E. & Brownell, C. A. (2003). Childhood peer relationships: social acceptance, friendships, and peer networks. *Journal of School Psychology*, *41*(4), 235–284. https://doi.org/10.1016/S0022-4405(03)00048-7

Händel, M., Vialle, W. & Ziegler, A. (2013). Student perceptions of high-achieving classmates. *High Ability Studies*, *24*(2), 99–114. https://doi.org/10.1080/13598139.2013.843139

Hannover, B. & Kessels, U. (2004). Self-to-prototype matching as a strategy for making academic choices. Why high school students do not like math and science. *Learning and Instruction*, *14*(1), 51–67. https://doi.org/10.1016/j.learninstruc.2003.10.002

Kessels, U. (2005). Fitting into the stereotype: How gender-stereotyped perceptions of prototypic peers relate to liking for school subjects. *European Journal of Psychology of Education*, *20*(3), 309–323. Verfügbar unter: http://www.jstor.org/stable/23421531

Kessels, U., Heyder, A., Latsch, M. & Hannover, B. (2014). How gender differences in academic engagement relate to students' gender identity. *Educational Research*, *56*(2), 220–229. https://doi.org/10.1080/00131881.2014.898916

Koster, M., Nakken, H., Pijl, S. J. & van Houten, E. (2009). Being part of the peer group: a literature study focusing on the social dimension of inclusion in education. *International Journal of Inclusive Education*, *13*(2), 117–140. https://doi.org/10.1080/13603110701284680

Kruse, H. & Kroneberg, C. (2020). *Contextualizing oppositional cultures: A multilevel network analysis of status orders in schools* (ECONtribute Discussion Papers Series 44). University of Bonn and University of Cologne, Germany. Verfügbar unter: https://ideas.repec.org/p/ajk/ajkdps/044.html

Ladd, G. W. (2007). Social learning in the peer context. In *Contemporary perspectives on socialization and social development in early childhood education* (Contemporary perspectives in early childhood education, pp. 133–164). Charlotte, NC: IAP.

Ladd, G. W., Kochenderfer, B. J. & Coleman, C. C. (1997). Classroom peer acceptance, friendship, and victimization: distinct relational systems that contribute uniquely to children's school adjustment? *Child Development*, *68*(6), 1181–1197. https://doi.org/10.1111/j.1467-8624.1997.tb01993.x

Osterman, K. F. (2000). Students' Need for Belonging in the School Community. *Review of Educational Research*, *70*(3), 323–367. Verfügbar unter: https://www.jstor.org/stable/1170786

Pelkner, A.-K. & Boehnke, K. (2003). Streber als Leistungsverweigerer? Projektidee und erstes Datenmaterial einer Studie zu mathematischen Schulleistungen [Nerds as refusers of performance?]. *Zeitschrift für Erziehungswissenschaft*, *6*(1), 106–125.

Pugh, M. J. & Hart, D. (1999). Identity development and peer group participation. *New Directions for Child and Adolescent Development*, *1999*(84), 55–70. https://doi.org/10.1002/cd.23219998406

Rentzsch, K., Schröder–Abé, M. & Schütz, A. (2013). Being Called A ‘Streber’: The Roles of Personality and Competition in the Labelling of Academically Oriented Students. *European Journal of Personality*, *27*(5), 411–423. https://doi.org/10.1002/per.1884

Turner, J. C. & Reynolds, K. J. (2012). Self-Categorization Theory. In P. van Lange, A. Kruglanski, E. Higgins & P. A. van Lange (Eds.), *Handbook of theories of social psychology* (vol. 2, S. 399–417). Los Angeles, Calif.: Sage. https://doi.org/10.4135/9781446249222.n46

Verhoeven, M., Poorthuis, A. M. G. & Volman, M. (2019). The Role of School in Adolescents’ Identity Development. A Literature Review. *Educational Psychology Review*, *31*(1), 35–63. https://doi.org/10.1007/s10648-018-9457-3

Wentzel, K. R., Jablansky, S. & Scalise, N. R. (2021). Peer social acceptance and academic achievement: A meta-analytic study. *Journal of Educational Psychology*, *113*(1), 157–180. https://doi.org/10.1037/edu0000468

Workman, J. & Heyder, A. (2020). Gender achievement gaps: the role of social costs to trying hard in high school. *Social Psychology of Education*, *23*(6), 1407–1427. https://doi.org/10.1007/s11218-020-09588-6

Youniss, J. & Haynie, D. L. (1992). Friendship in adolescence. *Developmental and Behavioral Pediatrics*, *13*(1), 59–66.

Zander, L., Kreutzmann, M. & Hannover, B. (2017). Peerbeziehungen im Klassenzimmer. *Zeitschrift für Erziehungswissenschaft* [Peer relations in the classroom], *20*(3), 353–386. https://doi.org/10.1007/s11618-017-0768-9

5

Gesamtdiskussion und Ausblick

1. Gesamtdiskussion und Ausblick

1.1. Zusammenfassung und Diskussion der Ergebnisse

1.1.1. Beitrag zur Definition von Leistungsstärke

Die vorliegende Promotionsarbeit beschäftigte sich mit leistungsstarken Schülerinnen und Schülern. Da bisher keine eigene Forschungstradition existiert, die sich spezifisch mit leistungsstarken Schülerinnen und Schülern beschäftigt und damit den Fokus weg vom Potenzial hin zur Performanz legt, war zunächst eine Eingrenzung des Begriffs wichtig. Hierfür wurden zunächst Begriffe von Leistung, die sich in Begabungs- und Expertiseforschung sowie in unterschiedlichen Forschungsbereichen, wie der pädagogischen Psychologie, Motivationspsychologie, Kognitionspsychologie und Bildungssoziologie, entwickelt haben, kontrastiert. In einem zweiten Schritt wurde ein Augenmerk darauf gelegt, wie leistungsstarke Schülerinnen und Schüler in unterschiedlichen Kontexten, nämlich der Bildungspolitik, der schulischen Praxis und der Forschung, identifiziert werden. Es wurde ein systematisches Review angefertigt, um einen Überblick über aktuell in der Forschung genutzte Operationalisierungen von Leistungsstärke zu gewinnen.

Der Überblick über den Umgang Forschender mit dem Begriff der Leistungsstärke veranlasste meine Koautoren und mich, Empfehlungen für die künftige Forschung zu Leistungsstarken abzugeben. Diese bezogen sich auf eine größere Transparenz der theoretischen Herleitung, der methodischen Darstellung und der Diskussion von Operationalisierungen bei Forschungsarbeiten zu leistungsstarken Schülerinnen und Schülern, um deren Nachvollziehbarkeit und Vergleichbarkeit, damit deren Anschlussfähigkeit an andere Forschungsarbeiten und deren potenzielle Generalisierbarkeit zu erhöhen. In Studie I wird die Aussage getroffen, dass eine stärkere Einheitlichkeit über Forschungsarbeiten hinweg ein anzustrebendes Ziel sein könnte. Hierbei gingen wir insbesondere darauf ein, dass die relative Größe der als leistungsstark bezeichneten Gruppe (gegenüber der Gesamtpopulation) ein Element sein könnte, welches über verschiedene Indikatoren hinweg ein gemeinsames Verständnis des Begriffs „leistungsstarke Schülerinnen und Schüler“ stiften könnte. Diese Anregung bietet durchaus Diskussionspotenzial. So wurde kritisch gesehen, ob ein solches gemeinsames Verständnis über verschiedene Forschungsbereiche hinweg überhaupt notwendig sei. Allerdings kann zumindest ein innerhalb von Forschungsfeldern geteiltes Verständnis diese voranbringen, indem Forschungsarbeiten besser aufeinander beziehbar sind und aufeinander aufgebaut werden können. Wissenschaft benötigt klare Begriffe und Definitionen. Wenn man beginnt, diese Definitionen explizit zu machen und konsequent anzuwenden, wird der kumulative Aufbau von Wissen beschleunigt. Die

Verwendung einer Vielzahl nicht klar definierter Terme ist ein generelles Problem in der psychologischen Forschung, welches in vielen Bereichen unter dem Begriff „jingle-jangle fallacy“ kritisiert wurde (z. B. Persönlichkeit: Block, 1995, Motivationsforschung: Marsh et al., 2019 Marsh, Craven, Hinkley & Debus, 2003). Auch in bisherigen Arbeiten mit Bezug zu Leistungsstärke wurden eine Vielzahl von Begriffen verwendet: „Leistungsexzellenz“, „Hochleistung“, „hohe Kompetenz“, “Leistungsstärke“, „Spitzenleistung“ aber auch „top performers“, „high bzw. superior performers“ oder „high achievers“. Dies bietet ggf. auch die Chance, unterschiedliche Definitionen voneinander abzugrenzen, die parallel in Verwendung sein können. So wird in der Marburger Hochbegabtenstudie von „Hochleistern“ gesprochen und eine Parallelität mit den Hochbegabten angestrebt (Rost, 2009). Das hier verwendete Kriterium von Leistungsstärke könnte also extremer sein als beispielsweise im Kontext der Large-Scale-Assessments, in denen gelegentlich von „top performern“ (Kompetenzstufe 5 und 6) die Sprache ist (z. B. Organisation for Economic Co-operation and Development [OECD], 2009, 2014, 2015).

Allerdings stellt sich auch die Frage, in welchen Fällen leistungsstarke Schülerinnen und Schüler überhaupt als separate Gruppe untersucht werden müssen und in welchen Fällen die Verwendung eines kontinuierlichen Leistungsmaßes eine gute Alternative darstellt. Die Nutzung von Leistungsgruppen hat den Vorteil, dass Ergebnisse anschaulicher dargestellt werden und eine Visualisierung und Kommunikation der Befunde einfacher sein kann. Daneben werden in manchen Fällen Maße verwendet, deren Skalenniveau keine metrische Interpretation erlaubt (z.B. Schulform- oder Kurszugehörigkeit, Teilnahme an einem Förderprogramm, Kompetenzstufen, teils auch Schulnoten), da es sich um nominal- oder ordinalskalierte Merkmale handelt und auch das zugrundeliegende latente Konstrukt nicht normalverteilt ist. Hingegen wurde das Vorgehen der künstlichen Dichotomisierung metrischer Merkmale von einigen Forschenden aus statistischen Gesichtspunkten stark kritisiert. Insbesondere gehe die Dichotomisierung mit einem Informationsverlust einher, da graduelle Unterschiede in der Ausprägung der künstlich dichotomisierten Variablen verlorengehen. Da die Gruppenbildung zu einer geringeren Varianz führt, gehe die Dichotomisierung in der Regel mit einer Verringerung der Teststärke und der Reduktion der Effektgröße einher (DeCoster, Iselin & Gallucci, 2009; MacCallum, Zhang, Preacher & Rucker, 2002).

Als Ergebnis des systematischen Reviews wurde vorgeschlagen, das Potenzial von Multiversumsanalysen zu nutzen, um eine Vergleichbarkeit von Ergebnissen unterschiedlicher Studien zu ermöglichen, divergierende Ergebnisse besser verstehen zu können und zugleich die Robustheit von Befunden über verschiedene Operationalisierungen von Leistungsstärke hinweg zu prüfen. Dieser Ansatz wurden in den beiden weiteren Beiträgen verfolgt. Im ersten Beitrag

konnte so die Robustheit des Effekts demonstriert werden, während die Variation von Operationalisierungen im zweiten Beitrag einen Hinweis zu möglichen Ursachen divergierender Ergebnisse früherer Studien aus der Entwicklungspsychologie und der Sozialpsychologie lieferte.

1.1.2. Beitrag zu Matthäuseffekten

Der zweite Artikel der Dissertation beschäftigte sich mit der Frage, ob bei Schülerinnen und Schülern an Gymnasien von der fünften bis zur neunten Klassenstufe kumulative Vorteils- bzw. Matthäuseffekte beobachtbar sind oder eher kompensatorische Prozesse. Die Ergebnisse der Wachstumskurvenanalyse deuteten auf einen Kompensationseffekt hin, das heißt, Unterschiede zwischen leistungsstärkeren Schülerinnen und Schülern und ihren Mitschülerinnen und Mitschülern verringerten sich über die Zeit hinweg. Insbesondere im ersten Zeitabschnitt, dem Jahr nach dem Übertritt an das Gymnasium waren die Wachstumsraten bei Leistungsstärkeren, sowohl in Mathematik als auch im Lesen, geringer.

Als Ursachen dieses Befundes kommen unterschiedliche mögliche Interpretationen in Betracht. Dabei spielen sowohl methodische als auch inhaltliche Aspekte eine Rolle. Zunächst können kompensatorische Befunde entstehen, wenn ein Testinstrument eine eher geringe Reliabilität hat oder Boden- und Deckeneffekte aufweist (Pfost, Hattie, Dörfler & Artelt, 2014). In der vorliegenden Untersuchung wurden Tests eingesetzt, die den großen Vorteil hatten, mittels Ankeritemdesign längsschnittlich verknüpfbar und damit besonders geeignet für den gewählten Ansatz der Wachstumskurvenanalyse zu sein. Auch war die Reliabilität in beiden Domänen, Mathematik und Lesen, akzeptabel. Allerdings differenzierten die Tests im unteren Bereich besser und teilweise existierten Deckeneffekte im Lesetest. Dies stellt die Verlässlichkeit der Befunde für dieses Fach in Frage. Neben Aspekten des Testinstruments ist auch die Modellierung von Matthäuseffekten mit Schwierigkeiten behaftet. Ein grundsätzliches Problem bei der Analyse von Extremgruppen stellt das Phänomen der Regression zur Mitte dar, welches darin besteht, dass bei längsschnittlichen Erhebungen Messwerte von Personen, welche zum ersten Messzeitpunkt extremer waren, bei Folgeerhebungen generell näher am Mittelwert liegen. Damit ist die Wahrscheinlichkeit höher, kompensatorische Entwicklungen vorzufinden, wenn Gruppeneinteilungen zum ersten Messzeitpunkt vorgenommen werden.

Die Ergebnisse von Studie II sind aus diesen Gründen mit Vorsicht zu interpretieren. Nichtsdestotrotz reihen sich die Befunde in eine Reihe vorhergehender Studien ein, welche ebenfalls Kompensationseffekte fanden und können damit als erste Evidenz für Kompensationseffekte auch innerhalb dieser Altersgruppe und spezifisch an Gymnasien gewertet werden. Um die Resultate jedoch zu erhärten, wären weitere Studien nötig, welche eine längsschnittliche Erhebung Messinstrumenten, die auch im oberen Fähigkeitsbereich gut

differenzieren, umfassen. Idealerweise würde eine größere Zahl an Messzeitpunkten in dichteren Abständen eingeführt, um nichtlineare Verläufe engmaschiger darstellen zu können. Weiterhin könnten unterschiedliche Analysemodelle gegenübergestellt werden, welche sich eignen, längsschnittliche Entwicklungen festzuhalten. Neben Multigruppenmodellen können auch Interaktionsanalysen Aufschluss über Unterschiede je nach Leistungsniveau geben (DeCoster et al., 2009; MacCallum et al., 2002). Weiterhin sind auch Mehrebenenmodelle geeignet, um die zeitliche Entwicklung darzustellen und gleichzeitig Effekte innerhalb und zwischen Klassen zu differenzieren. Damit wäre eine Fortführung und Erweiterung der Forschung im Rahmen des Optimalklassenparadigmas (Schwippert, 2001) möglich, welche Klassen identifiziert, welche eine überdurchschnittliche Entwicklung mit Verringerung der Leistungsstreuung verbinden. Insbesondere wäre dann interessant, ob es möglich ist, Faktoren zu identifizieren, welche für differenzielle Wachstumsraten zwischen Leistungsstarken und ihren Mitschülerinnen und Mitschülern verantwortlich sind. Vielversprechend scheint dabei insbesondere, auch die sich ändernde Rolle von Unterrichtsmerkmalen über den Kompetenzerwerb hinweg in den Blick zu nehmen. So könnten beispielsweise direkt nach dem Schulübergang oder in bestimmten Stufen der Kompetenzentwicklung andere Instruktionsmethoden und Unterrichtscharakteristika bedeutsam sein als im weiteren Verlauf. Diese Perspektive ist auch neueren Talententwicklungsmodellen, wie dem Talent Development in Achievement Domains (TAD-Framework, Preckel et al., 2020) inhärent.

1.1.3. Beitrag zur sozialen Integration Leistungsstarker

Im dritten Teil der Dissertation wurde die soziale Integration leistungsstarker Schülerinnen und Schüler näher untersucht. Dabei wurde zunächst festgestellt, dass ein insgesamt positiver Zusammenhang zwischen Leistung und sozialer Integration in ihre Schulklasse, welcher bereits in vorhergehenden Untersuchungen festgestellt wurde, auch über unterschiedliche Facetten sozialer Integration und verschiedene Operationalisierungen von Leistung im Großen und Ganzen bestätigt werden konnte. Detailliertere Analysen zeigten weiterhin, dass die nichtlinearen Effekte sich bei den verschiedenen Facetten sozialer Integration unterschiedlich darstellten. So waren für psychometrische Maße lineare Zusammenhänge vorhanden, während für die soziometrischen Maße quadratische Trends nachzuweisen waren: die Einbindung in instrumentelle Netzwerke (Hilfesuche) zeigte einen deutlichen positiven quadratischen Trend, die übrigen Facetten (Freundschaft, Ablehnung und Pausenkontakte) hingegen einen umgekehrt quadratischen Trend, der darin bestand, dass der Zusammenhang im oberen Bereich der Leistungsstärke ein Plateau bildete. Weitergehend wurde der Zusammenhang zwischen geschlechtstypischen Leistungsprofilen, ihrer Ausprägung bei Jungen und Mädchen und der entsprechende

Zusammenhang zur sozialen Integration untersucht. Die erwarteten Interaktionen (Leistungsstärke in männlich bzw. weiblich konnotierten Fächern sollte je nach Geschlecht einen unterschiedlichen Zusammenhang mit sozialer Integration zeigen) waren jedoch allesamt nicht von Null verschieden.

Zunächst ist dieser Befund überraschend, da er mit Theorien, welche sich mit Intragruppenprozessen im Allgemeinen (Festinger, 1954; Schachter, 1952) und mit der Wirkung von stereotypen Wahrnehmungen von Schülerinnen und Schülern im Speziellen befassen (Kessels, Heyder, Latsch & Hannover, 2014; Mayeux & Kleiser, 2020), nicht im Einklang steht. Diesen Theorien und Befunden entsprechend wäre anzunehmen, dass Schülerinnen und Schüler, welche nicht dem vorherrschenden Geschlechtsstereotyp entsprechen und aus der Gruppe herausstechen, eher gemieden werden. Der entscheidende Unterschied zwischen der hier durchgeführten Studie und denjenigen, auf die sich Forschung zu Geschlechterstereotypen stützen ist, dass hier ein naturalistischer Kontext vorliegt, in welchem die Beliebtheitsratings realer Klassenkameraden und -kameradinnen vorgenommen wurden. Bisherige Forschungsarbeiten nutzten meist Vignetten von hypothetischen Leistungsstarken (z. B. Händel, Vialle & Ziegler, 2013) oder wahrgenommene injunktive Normen (Cialdini, Kallgren & Reno, 1991). Tatsächliches Sozialverhalten stand damit bisher nicht im Fokus dieser Untersuchungen.

Auch wenn injunktive Normen (die mit den zuvor genannten Verfahren eher gemessen werden) die Macht haben können, Verhalten zu steuern, so geschieht dies doch nur unter bestimmten Bedingungen (Kallgren, Reno & Cialdini, 2000). In der schulischen Realität ist die Fachleistung von Schülerinnen und Schülern lediglich eine Eigenschaft unter vielen anderen personalen Eigenschaften, welche einen potenziell größeren Einfluss auf soziale Interaktionen haben (man denke nur an soziale Kompetenz oder Persönlichkeit). Darüber hinaus gibt es Untersuchungen, die Unterschiede im Leistungsklima von Schulklassen (Boor-Klip, Segers, Hendrickx & Cillessen, 2017; Kruse & Kroneberg, 2020; Palacios et al., 2019; van Houtte, 2006) belegen, welche möglicherweise den Zusammenhang von Leistung und sozialer Integration moderieren. Auch die Rolle der Lehrkraft für das soziale Klima in einer Klasse wird diskutiert (Decristan, Kunter & Fauth, 2022; Farmer, McAuliffe Lines & Hamm, 2011). Um den Effekt des Klassenkontextes genauer zu untersuchen und von Einflüssen auf Ebene des Individuums zu trennen, wäre es eine Möglichkeit, sich anzuschauen, wie Schülerinnen und Schüler vor und nach einem Schulwechsel in ihre Klasse integriert sind – bei diesem Übergang bleiben die personalen Eigenschaften weitgehend konstant, während die Umwelt sich ändert. Allerdings müsste man sich bei einem solchen Untersuchungsdesign Gedanken darüber machen, dass die Beobachtungen, das heißt die soziale Integration der einzelnen der Schülerinnen und Schüler nicht unabhängig voneinander sind – eine Annahme, welche bei Standard-Analysemethoden meist notwendig ist. Spezielle

Techniken, wie beispielsweise dyadische Interdependenz-Modelle (Kenny, 1996) oder die Modelle der sozialen Netzwerkanalyse können mit diesem Problem umgehen, so dass ihre Anwendung in derartigen Fragestellungen in Betracht gezogen werden sollte.

Ein anderes gewichtiges Argument bei der Untersuchung des Phänomens ist, dass dadurch, dass der Zusammenhang zwischen gezeigter Leistung und sozialer Integration untersucht wurde, Klassen, in denen – zum Beispiel aufgrund oppositioneller Peernormen – Schülerinnen und Schüler sich gar nicht erst trauen, ihre Fähigkeiten zu zeigen oder gar zu entwickeln, gewissermaßen durch das Raster fallen (z. B. Pelkner, Günther & Boehnke, 2002). Auch qualitative Forschung zeigt eindrücklich, dass es begabte Underachiever gibt, bei denen der Wunsch nach Zugehörigkeit zu einer Peergruppe dazu führt, dass sie ihre tatsächlichen Fähigkeiten nicht zeigen (Clasen & Clasen, 1995). Tatsächlich war auch in unseren Multiversumsanalysen der Zusammenhang zwischen sozialer Integration und Kompetenzen im Mittel stärker als der Zusammenhang zwischen sozialer Integration und Noten, was darauf hinweisen könnte, dass Jugendliche mit hohen Fähigkeiten diese nicht vollständig im Unterricht demonstrieren, um bei Peers anerkannter zu sein. Diese Interpretation der Beobachtungen stellt allerdings einen großen Sprung dar und wäre durch die Entwicklung spezifischer Hypothesen und darauf abgestimmter Analysen zu überprüfen. Damit stellt sich weiterführend auch die Frage, ob die schulische Leistung lediglich indirekt mit sozialer Integration zusammenhängt, zum Beispiel durch gemeinsame Ursachen (Intelligenz als Prädiktor von Noten und sozialer Kompetenz). Hier könnten Pfadanalysen möglicherweise Aufschluss geben.

1.1.4. Methodischer Beitrag: Anwendung der Multiversumsanalyse

Anknüpfend an den ersten Artikel dieser Dissertation wurde die Anwendung der Multiversumsanalyse in unterschiedlichen Kontexten demonstriert. Dabei wurde zunächst das Potenzial gezeigt, die Generalisierbarkeit von Befunden zu prüfen, welches der Zweck ist, der meist mit der Anwendung von Multiversumsanalysen verbunden wird (Gelman & Loken, 2013; Gelman & Loken, 2014; Jansen, Neuendorf & Kocaj, 2021; Simonsohn, Simmons & Nelson, 2015). So stand im zweiten Artikel der Dissertation die Robustheitsprüfung im Mittelpunkt der Multiversumsanalyse. Weitergehend soll hier aber auch reflektiert werden, welche Rolle die Multiversumsanalyse als explorative Methode in der Entwicklung von Theorien haben kann. Dies geschieht durch Diskussion des dritten Beitrags, in welchem der Entdeckungszusammenhang einen Schwerpunkt bildete.

Motiv 1: Konfirmatorische Multiversumsanalyse als Robustheitsprüfung

Dieses Motiv wurde im zweiten Artikel der Dissertation aufgegriffen, indem die Ergebnisse der Untersuchung der Leistungsentwicklung leistungsstarker Schülerinnen und Schüler einem Robustheitscheck unterworfen wurden. Eine Dimension, auf welcher die Robustheit bzw. Varianz der interessierenden Effekte unter anderem untersucht wurde, stellt die Definition von Leistungsstärke dar. Speziell wurde neben dem statistischen Modell zur Berechnung der Leistungswerte auch untersucht, ob die Variation des Anteils von Schülerinnen und Schülern, die als leistungsstark charakterisiert werden, einen großen Einfluss auf die Ergebnisse hat. Es zeigte sich, dass der Unterschied in den Wachstumsraten zwischen den ersten beiden Messzeitpunkten geringer war, je größer die Subgruppe war, die als leistungsstark gruppiert wurden und je niedriger damit der Cut-off-Wert angesetzt wurde. Dennoch blieb der Unterschied in den Wachstumsraten zwischen den Gruppen auch deutlich, wenn ein Anteil von bis zu 50 Prozent als leistungsstark gruppiert wurde. Die Multiversumsanalyse zeigte damit, dass die grundsätzliche Aussage der Analysen sich über verschieden strenge Kriterien, die man an die Auswahl einer leistungsstarken Subgruppe anlegt, nicht änderte. Damit wurde der Befund als robust eingeschätzt.

Im dritten Artikel wurde mit der Variation der Operationalisierung von Leistungsstärke ebenfalls die Robustheit von Effekten abgeschätzt und weiterhin die Vergleichbarkeit mit anderen Studien erhöht. Denn in vergangenen Arbeiten gab es, was Leistung betrifft, unterschiedliche Herangehensweisen, diese zu operationalisieren – manchmal über die Note, über Selbsteinschätzungen oder Fremdeinschätzungen. Daher wurde hier die Frage in den Blick genommen, wie Aussagen über die soziale Integration in Abhängigkeit von der Operationalisierung und Definition von Leistungsstärke variierten. Dabei wurden sowohl der Indikator als auch der Fachbezug und die Bezugsnorm der Operationalisierung von Leistungsstärke variiert und die jeweiligen Zusammenhänge mit verschiedenen Indikatoren sozialer Integration dargestellt. Es zeigte sich, dass die Zusammenhänge insgesamt schwach negativ bis leicht positiv waren, sich aber auch systematische Unterschiede zwischen den verschiedenen Operationalisierungen zeigten.

Damit wurde in der vorliegenden Dissertation zunächst vorgeschlagen, dann aber auch demonstriert, wie Large-Scale-Assessments in Verbindung mit Multiversumsanalysen genutzt werden können, um die Effekte unterschiedlicher Operationalisierungen von Leistungsstärke gegenüberzustellen und damit die Robustheit der Ergebnisse zu prüfen. Es konnte jeweils gezeigt werden, dass es systematische Unterschiede zwischen verschiedenen Leistungsindikatoren gab, und in welcher Größenordnung der Einfluss der Operationalisierung auf die Ergebnisse lag.

Motiv 2: Exploratorische Multiversumsanalyse zur Theorieentwicklung

Im dritten Artikel dieser Arbeit wurde die Multiversumsanalyse aber auch als Werkzeug eingesetzt, um das soziale Erleben leistungsstarker Schülerinnen und Schüler in unterschiedlichen Facetten zu beleuchten. Bisherige Studien nutzten sehr unterschiedliche Methoden, soziale Integration zu erfassen. Manche Studien verwendeten verschiedene soziometrische Maße, andere nutzten psychometrische Maße von Eingebundenheit, wieder andere nutzten Vignetten, um den Effekt von Leistung auf soziale Eingebundenheit abzuschätzen oder aber qualitative Befragungen. Diese Vielfalt an Operationalisierungen wurde allerdings nicht reflektiert oder in einem übergreifenden Framework von sozialer Integration verortet. Dies ist ein Hinweis darauf, dass das Konzept sozialer Integration bisher noch nicht in einem Maße ausdifferenziert wurde, um Grundlage einer ausformulierten Theorie der sozialen Integration bilden zu können (Koster, Nakken, Pijl & van Houten, 2009). Mit der Anwendung der Multiversumsanalyse in diesem Kontext wurden Anfänge einer Theoriebildung fortgeführt, indem aufbauend auf Ergebnissen von Literaturreviews (Bossaert, Colpin, Pijl & Petry, 2013; Koster et al., 2009) die Konzeptualisierungen weiter anhand empirischer Daten exploriert wurden. Dabei wurde eine Reihe von Mustern entdeckt: Unter Nutzung von Kompetenzen zeigten sich generell höhere Zusammenhänge zu sozialer Integration. Das akademische Selbstkonzept war insgesamt schwächer mit sozialer Integration assoziiert, in einigen Fällen (Facette Freundschaft) waren die Zusammenhänge gar negativ. Lediglich für die Facette der subjektiven Integration war die Korrelation mit dem akademischem Selbstkonzept höher.

Die Fächer betreffend zeigte sich, dass die Deutschnote einen stärkeren Zusammenhang mit Freundschaften hatte, während bei den anderen Facetten sozialer Integration das generelle Leistungsniveau, d.h. Mittelwerte, einen stärkeren Zusammenhang mit Integration zeigten.

Betrachtet man die Bezugsnorm, zeigte sich, dass die stärksten Korrelationen bestanden, wenn rohe Noten- bzw. Leistungswerte genutzt wurden. Etwas niedriger fielen die Zusammenhänge bei Verwendung klassenzentrierter Werte aus, die geringsten und teils sogar negative Zusammenhänge zeigten sich zwischen sozialer Integration und Leistungsstärke, wenn lediglich Klassen einbezogen wurden, in denen nur eine einzige Person die Bestnote erhielt.

Hier zeigt sich, dass das Zusammenspiel unterschiedlicher Dimensionen der Operationalisierung von Leistungsstärke (und auch unterschiedlicher Facetten sozialer Integration) einen ganzen Raum von Fragen öffnet und zur Entwicklung von Theorien und Hypothesen einlädt: Werden Zusammenhänge zwischen Leistung und Integration über soziale Kompetenzen vermittelt (die möglicherweise in verschiedenen Fächern in unterschiedlichem Maße Eingang in die Note finden)? Sind sprachliche Kompetenzen eine besondere Voraussetzung für die Bildung von

Freundschaften? Durch welche Merkmale auf Personenebene lassen sich Unterschiede zwischen subjektiver und soziometrischer Integration erklären? Gibt es verschiedene Gruppen von Schülerinnen und Schülern, für die unterschiedliche Formen bzw. Facetten der Integration zu Zufriedenheit mit dem Sozialleben führt? In welchen Fällen führt die „Andersartigkeit“, das heißt, das Herausstechen aus der Leistungsverteilung in der Klasse zu negativen Effekten auf die soziale Integration? Welche Rahmenbedingungen sind förderlich für die sozialen Beziehungen einer Schulklasse? Diese und viele weitere Anschlussfragen könnten sich aus den vorgestellten Analysen ergeben. Dies demonstriert das Potenzial, welches dem Einsatz von Multiversumsanalysen und der Darstellung als Spezifikationskurve im Rahmen explorativer Analysen innewohnt. In dieser Phase, auch Entdeckungszusammenhang (Hoyningen-Huene, 1987) genannt, stehen die Operationalisierungen nicht, wie in den Artikeln von Steegen, Tuerlinckx, Gelman und Vanpaemel (2016) und Simonsohn et al. (2015, 2020) als prinzipiell austauschbare Entscheidungen im Datenaufbereitungs- und Analyseprozess nebeneinander, deren Variation eine Aussage über die Robustheit oder die tatsächliche Stärke eines Effekts erlauben soll. Stattdessen werden bewusst verschiedene Entscheidungen gegenübergestellt und kontrastiert, um Systematiken zu erkennen und auffällige Befunde in weiteren Untersuchungen gezielt replizieren und durch Theoriebildung erklären zu können. Damit wird hier dafür plädiert, den Aspekt der Weiterentwicklung von Theorien, der bei Steegen et al. (2016) in Ansätzen besprochen wird, zu stärken und die Methode noch bewusster in diesem Sinne zu nutzen.

In der aktuellen Forschung steht sehr häufig der Begründungszusammenhang im Fokus, das bedeutet, dass Forschende viel Zeit und Energie auf das Testen von Hypothesen verwenden und das Entwickeln von Hypothesen und Theorien weniger Aufmerksamkeit erfährt (Scheel, Tiokhin, Isager & Lakens, 2020; Swedberg, 2012). Swedberg (2012) plädiert dafür, sich bei der Theorieentwicklung auf empirisches Material zu stützen, und aufbauend auf der intensiven Beschäftigung mit den Daten kreative Ideen zu entwickeln, um diese dann durch die Anwendung weiterer Techniken zu Theorien auszuformulieren. Hier wird vorgeschlagen, die Multiversumsanalyse ebenfalls als Instrument im Rahmen des Entdeckungszusammenhangs einzusetzen, indem die Arbeit mit dem Datenmaterial und das Entdecken von Mustern eine Inspirationsquelle für die Entwicklung von Theorien wird. Scheel et al. (2020) beschreiben den Schritt der Konzeptbildung als weiteren notwendigen Schritt bei der Theorieentwicklung, bei welchem zunächst kohärente, wohldefinierte Konstrukte entwickelt werden, die sich empirisch gegenüber anderen Begriffen abgrenzen lassen. Auch dieser Prozess kann durch die Nutzung von Multiversumsanalysen unterstützt werden, indem mehrere Konzepte nebeneinandergestellt und systematische Unterschiede in Zusammenhängen zu anderen Konstrukten entdeckt werden.

Grenzen der Multiversumsanalyse

Der Ansatz der Multiversumsanalyse unterliegt bestimmten Voraussetzungen. Das Vorgehen ist, verglichen mit anderen Ansätzen, relativ ressourcenintensiv. Zunächst hat die Datengrundlage einen Einfluss darauf, wie viele Freiheitsgrade bei der Operationalisierung zur Verfügung stehen: Ein Datensatz, der viele unterschiedliche Indikatoren und eine große Stichprobe enthält, ermöglicht eine vielfältigere Spezifikation der zentralen Konstrukte und möglicher Analysen. Damit eignen sich besonders Large-Scale-Assessments für die Durchführung von Multiversumsanalysen. Aber auch Ressourcen, was die Rechen- und Speicherkapazität angeht, können bei manchen Analysen zu Einschränkungen führen. Schließlich erfordert ein Multiversum an Ergebnissen, je nach Zielstellung der Analysen, in manchen Fällen eine detaillierte Beschreibung und Interpretation der Ergebnisse, welche in Zeitschriftenaufsätzen mitunter sehr verkürzt stattfinden muss.

Zusammenfassend hat die vorliegende Arbeit gezeigt, wie die Variation von Operationalisierungen von Leistung bzw. Leistungsstärke im Rahmen von Multiversumsanalysen zum einen die Robustheit von Befunden absichern kann. Ein zweiter Aspekt bei der Durchführung von Multiversumsanalysen ist aber die Weiterentwicklung von Theorien, wenn systematische Abweichungen zwischen den verschiedenen Operationalisierungen gefunden werden können.

1.2. Ausblick

In der vorliegenden Arbeit wurde gezeigt, dass allgemeingültige Aussagen über „leistungsstarke Schülerinnen und Schüler“ irreführend sein können, denn zumeist ist nicht klar, wer genau gemeint ist, wenn einfach von leistungsstarken Schülerinnen und Schülern gesprochen wird. Gerade in der öffentlichen Diskussion, aber auch in der Bildungspolitik sind mitunter ganz unterschiedliche Gruppen mit einem solchen Begriff gemeint. Auch in der Forschung gibt es bisher keine Konvention dazu. Wie schulische Leistungsstärke definiert und operationalisiert wird, hängt zwar unter anderem von der Fragestellung und dem Untersuchungsdesign ab, in den meisten Fällen gibt es aber auch gewisse Freiheitsgrade, die zunächst nicht durch die Theorie ableitbar sind. Dieser Ambiguität kann begegnet werden, indem multiple Operationalisierungen für Leistungsstärke, zum Beispiel im Rahmen von Multiversumsanalysen, nebeneinandergestellt werden. Damit wird deutlich gemacht, wie variabel Ergebnisse in Abhängigkeit von der Operationalisierung sind. Dies ist zum einen im Kontext der Überprüfung von Hypothesen, im Rahmen eines Robustheitschecks wichtig.

Ein zweiter Anwendungsfall für Multiversumsanalysen, welcher bisher nicht im Fokus der Forschung stand, ist die explorative Multiversumsanalyse, welche nicht austauschbare Entscheidungen nebeneinanderstellt, sondern bewusst unterschiedliche Definitionen des

Konstrukts miteinander kontrastiert, um so zur Integration eines Forschungsbereiches und zur nachfolgenden Theoriebildung einzuladen.

Künftig könnte der hier vorgestellte Ansatz beispielsweise im Rahmen des TAD-Frameworks genutzt werden, um zu differenzieren, in welchen Abschnitten des Talententwicklungsprozesses welche Teilkompetenzen besonders bedeutsam für das Meistern nächster Entwicklungsschritte sind, indem unterschiedliche Aspekte von Leistungsstärke nebeneinandergestellt und in ihrem Zusammenhang zur weiteren Leistungsentwicklung in verschiedenen Entwicklungsabschnitten verglichen werden. Darstellungen von Multiversumsanalysen wie die Spezifikationskurve könnten in diesem Zusammenhang auch eine Rolle bei der Kommunikation von Forschungsergebnissen spielen, indem simultan Effekte auf eine Variation von Zielindikatoren dargestellt werden.

Literaturverzeichnis

Block, J. (1995). A contrarian view of the Five-Factor Approach to personality description. *Psychological Bulletin*, *117*(2), 187–215. https://doi.org/10.1037/0033-2909.117.2.187

Boor-Klip, H. J., Segers, E., Hendrickx, M. M. H. G. & Cillessen, A. H. N. (2017). The Moderating Role of Classroom Descriptive Norms in the Association of Student Behavior With Social Preference and Popularity. *The Journal of Early Adolescence*, *37*(3), 387–413. https://doi.org/10.1177/0272431615609158

Bossaert, G., Colpin, H., Pijl, S. J. & Petry, K. (2013). Truly included? A literature study focusing on the social dimension of inclusion in education. *International Journal of Inclusive Education*, *17*(1), 60–79. https://doi.org/10.1080/13603116.2011.580464

Cialdini, R. B., Kallgren, C. A. & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In *Advances in Experimental Social Psychology* (S. 201–234). Elsevier. https://doi.org/10.1016/s0065-2601(08)60330-5

Clasen, D. R. & Clasen, R. E. (1995). Underachievement of Highly able Students and the Peer Society. *Gifted and Talented International*, *10*(2), 67–75. https://doi.org/10.1080/15332276.1995.11672824

DeCoster, J., Iselin, A.-M. R. & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, *14*(4), 349–366. https://doi.org/10.1037/a0016956

Decristan, J., Kunter, M. & Fauth, B. (2022). Die Bedeutung individueller Merkmale und konstruktiver Unterstützung der Lehrkraft für die soziale Integration von Schülerinnen und Schülern im Mathematikunterricht der Sekundarstufe. *Zeitschrift für Pädagogische Psychologie*, *36*(1-2), 85–100. https://doi.org/10.1024/1010-0652/a000329

Farmer, T. W., McAuliffe Lines, M. & Hamm, J. V. (2011). Revealing the invisible hand: The role of teachers in children's peer experiences. *Journal of Applied Developmental Psychology*, *32*(5), 247–256. https://doi.org/10.1016/j.appdev.2011.04.006

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Gelman, A. & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem*: Columbia University. Verfügbar unter: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A. & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, *102*(6), 460–465. Verfügbar unter: http://www.jstor.org/stable/43707868

Händel, M., Vialle, W. & Ziegler, A. (2013). Student perceptions of high-achieving classmates. *High Ability Studies*, *24*(2), 99–114. https://doi.org/10.1080/13598139.2013.843139

Hoyningen-Huene, P. (1987). Context of discovery and context of justification. *Studies in History and Philosophy of Science*, *18*, 501–515.

Jansen, M., Neuendorf, C. & Kocaj, A. (2021). Welche Potenziale bieten Sekundäranalysen für die Erhöhung von Forschungsqualität und Replizierbarkeit. *Zeitschrift für Pädagogik*, *67*(6), 840–859. https://doi.org/10.3262/ZP2106840

Kallgren, C. A., Reno, R. R. & Cialdini, R. B. (2000). A Focus Theory of Normative Conduct: When Norms Do and Do not Affect Behavior. *Personality & social psychology bulletin*, *26*(8), 1002–1012. https://doi.org/10.1177/01461672002610009

Kenny, D. A. (1996). Models of Non-Independence in Dyadic Research. *Journal of Social and Personal Relationships*, *13*(2), 279–294. https://doi.org/10.1177/0265407596132007

Kessels, U., Heyder, A., Latsch, M. & Hannover, B. (2014). How gender differences in academic engagement relate to students' gender identity. *Educational Research*, *56*(2), 220–229. https://doi.org/10.1080/00131881.2014.898916

Koster, M., Nakken, H., Pijl, S. J. & van Houten, E. (2009). Being part of the peer group: a literature study focusing on the social dimension of inclusion in education. *International Journal of Inclusive Education*, *13*(2), 117–140. https://doi.org/10.1080/13603110701284680

Kruse, H. & Kroneberg, C. (2020). *Contextualizing oppositional cultures: A multilevel network analysis of status orders in schools* (ECONtribute Discussion Papers Series 44). University of Bonn and University of Cologne, Germany. Verfügbar unter: https://ideas.repec.org/p/ajk/ajkdps/044.html

MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*(1), 19–40. https://doi.org/10.1037//1082-989X.7.1.19

Marsh, H. W., Craven, R. G., Hinkley, J. W. & Debus, R. L. (2003). Evaluation of the Big-Two-Factor Theory of Academic Motivation Orientations: An Evaluation of Jingle-Jangle Fallacies. *Multivariate behavioral research*, *38*(2), 189–224. https://doi.org/10.1207/S15327906MBR3802_3

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T. et al. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, *111*(2), 331–353. https://doi.org/10.1037/edu0000281

Mayeux, L. & Kleiser, M. (2020). A Gender Prototypicality Theory of Adolescent Peer Popularity. *Adolescent Research Review*, *5*(3), 295–306. https://doi.org/10.1007/s40894-019-00123-z

Organisation for Economic Co-operation and Development. (2009). *Top of the class. High performers in science in PISA 2006*. OECD Publishing. https://doi.org/10.1787/9789264060777-en

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 results: What students know and can do. Student performance in mathematics, reading and science* (Vol. I, Rev. ed., Febr. 2014). Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2015). *The ABC of Gender Equality in Education. Aptitude, behaviour, confidence*. OECD Publishing. https://doi.org/10.1787/9789264229945-en

Palacios, D., Dijkstra, J. K., Villalobos, C., Treviño, E., Berger, C., Huisman, M. et al. (2019). Classroom ability composition and the role of academic performance and school misconduct in the formation of academic and friendship networks. *Journal of School Psychology*, *74*, 58–73. https://doi.org/10.1016/j.jsp.2019.05.006

Pelkner, A.-K., Günther, R. & Boehnke, K. (2002). Die Angst vor sozialer Ausgrenzung als leistungshemmender Faktor? Zum Stellenwert guter mathematischer Schulleistungen unter Gleichaltrigen. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen.* Weinheim: Beltz. https://doi.org/10.25656/01:23587

Pfost, M., Hattie, J., Dörfler, T. & Artelt, C. (2014). Individual differences in reading development. A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research*, *84*(2), 203–244. https://doi.org/10.3102/0034654313509492

Preckel, F., Golle, J., Grabner, R., Jarvin, L., Kozbelt, A., Müllensiefen, D. et al. (2020). Talent Development in Achievement Domains: A Psychological Framework for Within- and Cross-Domain Research. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, 691–722. https://doi.org/10.1177/1745691619895030

Rost, D. H. (Hrsg.). (2009). *Hochbegabte und hochleistende Jugendliche. Befunde aus dem Marburger Hochbegabtenprojekt* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 72, 2., erw. Aufl.). Münster: Waxmann.

Schachter, S. (1952). Deviation, rejection and communication. In L. Festinger, K. Back, S. Schachter, H. H. Kelley & J. Thibaut (Hrsg.), *Theory and experiment in social communication* (S. 51–82). Ann Arbor, Michigan: Edwards Borthers, Inc.

Scheel, A. M., Tiokhin, L., Isager, P. M. & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, 1745691620966795. https://doi.org/10.1177/1745691620966795

Schwippert, K. (2001). *Optimalklassen: mehrebenenanalytische Untersuchungen. Eine Analyse hierarchisch strukturierter Daten am Beispiel des Leseverständnisses* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 27). Münster, München [u.a.]: Waxmann.

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Swedberg, R. (2012). Theorizing in sociology and social science: turning to the context of discovery. *Theory and Society*, *41*(1), 1–40. https://doi.org/10.1007/s11186-011-9161-5

Van Houtte, M. (2006). School type and academic culture: evidence for the differentiation–polarization theory. *Journal of Curriculum Studies*, *38*(3), 273–292. https://doi.org/10.1080/00220270500363661

Erklärung

Erklärung über eigene Beiträge bei eingereichten Gemeinschaftsveröffentlichungen (nach CRediT Contributor Roles Taxonomy, s. https://casrai.org/credit/)

Artikel I:

CN: Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Visualization, Writing - original draft, Writing – revision

MJ: Conceptualization, Critical Review, Writing – revision

PK: Conceptualization, Critical Review

MV: Critical review

Artikel II:

CN: Conceptualization, Formal Analysis, Methodology, Visualization, Writing original draft, Writing revision

MJ: Conceptualization, Critical review, Writing revision

PK: Conceptualization, Critical review

Artikel III:

CN: Conceptualization, Formal Analysis, Methodology, Visualization, Writing original draft, Writing revision

MJ: Conceptualization, Critical review

Publikationsübersicht

Aufsätze in Zeitschriften mit Peer Review

Neuendorf, C., Jansen, M., Kuhl, P. & Vock, M. (2022). Wer ist leistungsstark? Operationalisierung von Leistungsstärke in der empirischen Bildungsforschung seit dem Jahr 2000. *Zeitschrift für Pädagogische Psychologie*. Advance online publication. https://doi.org/10.1024/1010-0652/a000343

Jansen, M., **Neuendorf, C.** & Kocaj, A. (2021). Welche Potenziale bieten Sekundäranalysen für die Erhöhung von Forschungsqualität und Replizierbarkeit? Zur Rolle von Multiversumsanalysen und integrativen Datenanalysen für die Bestimmung der Robustheit und Generalisierbarkeit von Forschungsbefunden. *Zeitschrift für Pädagogik 6*, 840-859. https://doi.org/10.3262/ZP2106840

Pegelow, L., Jansen, M., & **Neuendorf, C.** (2021). Erwerb des Zertifikats CoreTrustSeal (CTS) durch ein Forschungsdatenzentrum im Bildungsbereich – Motivation, Umsetzung und Lessons learned. *Bausteine Forschungsdatenmanagement 1*, 10-21. https://doi.org/10.17192/bfdm.2021.1.8310

Neuendorf, C., Jansen, M. & Kuhl, P. (2020) Competence Development of High Achievers Within the Highest Track in German Secondary School: Evidence for Matthew Effects or Compensation? *Learning and Individual Differences 77*, 101816. https://doi.org/10.1016/j.lindif.2019.101816

Lee, S., Davis, K. D., **Neuendorf, C.**, Grandey, A., Lam, C. B. & Almeida, D. M. (2016). Individual- and Organization-Level Work-to-Family Spillover Are Uniquely Associated with Hotel Managers' Work Exhaustion and Satisfaction. *Frontiers in Psychology, 7*, 1180. http://doi.org/10.3389/fpsyg.2016.01180

Lee, B., Lawson, K.M., Chang, P.J., **Neuendorf, C.**, Dimitrieva, N., & Almeida, D.M. (2015). Leisure-time physical activity moderates the longitudinal associations between work-family spillover and physical health, *Journal of Leisure Research 47*(4), 444-466.

Kapitel in Herausgeberwerken

Neuendorf, C., Kuhl, P., & Jansen, M. (2017). Leistungsstarke Schülerinnen und Schüler in Deutschland. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich, & N. Haag (Hrsg.), IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich (S. 317-334). Münster: Waxmann.

Weitere Publikationen

Neuendorf, C., Kocaj, A., Rüdiger, C., & Jansen, M. (2022). Thematisierung von Replikationen und Open Science Praktiken in Lehre und Studium – Die Rolle von Sekundärdatenanalysen. *Psychologische Rundschau 73*(1). https://doi.org/10.1026/0033-3042/a000575

Neuendorf, C., Jansen, M., & Pegelow, L. (2020). Assessing the re-use potential of research data in empirical educational research. *RatSWD Working Paper Series 270*. http://doi.org/10.17620/02671.49

Pegelow, L., **Neuendorf, C.**, Daniel, A., Buck, D. (2020). Formulierungshilfen für Forschungsdatenzentren zum Thema Nutzungsbedingungen. *RatSWD Working Paper Series, 271*. https://doi.org/10.17620/02671.53

Verbund Forschungsdaten Bildung (2019): Hinweise zur Codierung fehlender Werte in der Aufbereitung quantitativer Daten. *fdbinfo Nr. 6*.

Neuendorf, C., & Jansen, M. (2018). Bericht vom Workshop „Nachnutzungspotenzial von Forschungsdaten“ des Verbunds Forschungsdaten Bildung. *forschungsdaten bildung informiert, Nr. 8.*

Meyermann, A., Bambey, D., Ebel, T., Eisentraut, M., Jansen, M., & Kuhl, P., … Trixa, J. (2017). Schlussbericht zum Verbundprojekt „Sicherung und Nachnutzung der Forschungsdaten des Rahmenprogramms zur Förderung der empirischen Bildungsforschung“: Projektlaufzeit: 01.10.2013 bis 30.09.2016. https://doi.org/10.2314/gbv:897124898

Meyermann, A., Bambey, D., Jansen, M., Mauer, R., Ebel, T., Eisentraut, M., Harzenetter, K., Kuhl, P., **Neuendorf, C.**, Pegelow, L., Porzelt, M., Rittberger, M., Schwager, T., Stanat, P. & Trixa, J. (2017). Der Verbund Forschungsdaten Bildung - Eine Forschungsdateninfrastruktur für die empirische Bildungsforschung. *RatSWD Working Paper Series 266*.