Schult, Johannes

VERA fair-bessern. Ein Vergleich von Strategien zur Berücksichtigung von Kontextbedingungen bei schulischen Vergleichsarbeiten

Journal for educational research online 14 (2022) 2, S. 3-24

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.02.01
Volume 14 (2022), No. 2, S. 3–24

Johannes Schult

VERA fair-bessern. Ein Vergleich von Strategien zur Berücksichtigung von Kontextbedingungen bei schulischen Vergleichsarbeiten

Zusammenfassung

Um spezifische Schul- und Unterrichtseffekte abzuschätzen, werden in Lernstandserhebungen wie VERA (VERgleichsArbeiten) neben dem Landesmittelwert häufig adjustierte Schulwerte für einen fairen Vergleich rückgemeldet. Dabei wird als Vergleichswert die Leistung ähnlicher Schulen berechnet, um Einflüsse zu berücksichtigen, die jenseits des Handlungsspielraums der Lehrkräfte und Schulleitungen liegen. Zu den Adjustierungsstrategien gehören die Vergleiche mit Subgruppenmittelwerten, mit ähnlichen existierenden Schulen und mit modellbasierten Erwartungswerten. Anhand zweier Vollerhebungen von VERA 3 (je über 2300 Schulen) und VERA 8 (je über 1200 Schulen) in Baden-Württemberg aus den Jahren 2019 und 2021 wurden die Adjustierungsstrategien hinsichtlich ihrer Fairness (Determinationskoeffizient R^2*) und Standardfehler miteinander verglichen. Kontextvariablen waren dabei Geschlecht, Alltagssprache, Migrationshintergrund, soziokulturelles Kapital sowie zusätzlich bei VERA 8 Schulart und Vorwissen. Die modellbasierten Erwartungswerte zeigten fächerübergreifend die größten* R^2*-Werte und die kleinsten Standardfehler. Dieser Fairness-Vorsprung geht dabei nicht zulasten der Testökonomie, denn die verwendeten Kontextvariablen werden für jede Adjustierungsstrategie benötigt. Die Ergebnisse zeigen zudem, dass die Heterogenität der Testleistungen stark vom sozialen Kontext abhängt. Vor diesem Hintergrund kann VERA mit der Rückmeldung des fairen Vergleichs ein hilfreicher Baustein für die datenbasierte Unterrichts- und Schulentwicklung sein.*

Schlagworte

fairer Vergleich, Vergleichsarbeiten, Adjustierung, Heterogenität

Dr. Johannes Schult, ORCID: 0000-0003-4576-4605, Institut für Bildungsanalysen Baden-Württemberg (IBBW), Heilbronner Straße 172, 70191 Stuttgart
E-Mail: Johannes.Schult@ibbw.kv.bwl.de

Johannes Schult

Improving Comparative Performance Tests: A Comparison of Procedures to Adjust Test Scores for School Context

Abstract

In order to estimate specific effects of schooling and instruction, state-wide proficiency tests like VERA (VERgleichsArbeiten) often report not only the state's mean score, but also an adjusted school score for a so-called fair comparison. The average performance of similar schools is estimated for the fair comparison with the aim of controlling for influences that lie outside the latitude of teachers and school management. Adjustment strategies include comparisons with subgroup mean scores, mean scores of similar existing schools, and model-based expected values. Using the large-scale assessments VERA 3 (over 2300 schools) and VERA 8 (over 1200 schools) in Baden-Württemberg from 2019 and 2021, adjustment strategies were compared with regard to their fairness (coefficient of determination R^2*) and standard errors. Variables for contextualization were gender, everyday language, migration background, socio-cultural capital, and, additionally in VERA 8, type of school and previous knowledge. Model-based expected values yielded the largest* R^2 *values and the smallest standard errors across all subjects. This advantage in terms of fairness is not to the disadvantage of the test efficiency, because the context variables are needed for every adjustment strategy. The results also show that the heterogeneity of test performances is considerably influenced by the social context. Against this backdrop, VERA featuring a fair comparison feedback can be a helpful element for school development.*

Keywords

fair comparisons, Vergleichsarbeiten, value-added models, heterogeneity

1. Einleitung

Die Vergleichsarbeiten VERA sind in Deutschland ein wichtiger Bestandteil der Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring (Kultusministerkonferenz [KMK], 2015). Sie sind ein Verfahren zur Qualitätssicherung auf Ebene der Schulen, welches die Unterrichts- und Schulentwicklung flächendeckend unterstützen soll. Eine zentrale Rolle spielen dabei die bildungsstandardorientierten Rückmeldungen, anhand derer eine Standortbestimmung im Hinblick auf die Landesergebnisse stattfinden kann. Die Landeswerte spiegeln jedoch nicht nur die im Unterricht erworbenen Kompetenzen wider, sondern auch eine Vielzahl an Einflüssen, die jenseits des Handlungsspielraums der Lehrkräfte und Schulleitungen liegen. Um spezifische Schul- und Unterrichtseffekte abzuschätzen, werden in vielen Bundesländern bei VERA auch adjustierte Werte berechnet und den Schulen zurückgemeldet (vgl. Tarkian, Maritzen, Eckert & Thiel, 2019). Dabei kommen je

nach Land und je nach Schuljahr unterschiedliche Adjustierungsverfahren zum Einsatz (Fiege, 2016). Die vorliegende Studie vergleicht anhand empirischer Vollerhebungen von VERA 3 und VERA 8 die Fairness der typischen Adjustierungen und setzt diese in Bezug zur jeweiligen Praktikabilität.

1.1 Adjustierungsstrategien für faire Vergleiche

Die Leistungen von Schüler:innen in Kompetenztests hängen von verschiedenen Einflüssen ab. Dazu gehören neben dem Unterricht auch individuelle Voraussetzungen wie Vorwissen und soziokulturelle Faktoren (Dumont et al., 2013). Die Zusammensetzung der Schülerschaft einer Schule beeinflusst daher auch die Ergebnisse, die im Rahmen der Vergleichsarbeiten VERA erzielt werden. Die unterschiedlichen Voraussetzungen können unter anderem bei einem sogenannten *fairen Vergleich* berücksichtigt werden, der über die alleinige Betrachtung einer Verortung der eigenen Ergebnisse im Landesvergleich hinaus Impulse für die Schul- und Unterrichtsentwicklung liefert (Fiege, Reuther & Nachtigall, 2011)[1]. Der faire Vergleich entsteht durch die Adjustierung der Vergleichswerte mithilfe von bestimmten Kontextvariablen. Da die Kausalstruktur der Einflussfaktoren nie vollständig erfasst und modelliert werden kann, können in der Praxis lediglich „*fairere* Vergleiche ermöglicht werden“ (Fiege, 2016, S. 92). Für die Adjustierung gibt es mehrere Strategien (vgl. Fiege, 2014, S. 72).

1.1.1 Vergleich mit dem Landesmittelwert

Der Landesmittelwert ergibt sich aus der mittleren Leistung aller teilnehmenden Schüler:innen eines Landes. Bei VERA handelt es sich dabei üblicherweise um ein Bundesland. Einen Bundesmittelwert gibt es aufgrund der länderspezifischen Durchführungen und der formalen Rahmenbedingungen, die Bundeslandvergleiche verbieten, nicht. Lehrkräfte und Schulleitungen können die Leistung ihrer Klasse bzw. Schule im Rahmen der sozialen Vergleichsnorm hinsichtlich der Leistung der Gesamtheit aller teilnehmenden Schulen abgleichen. In den Landesmittelwert fließen sowohl schulische als auch außerschulische Faktoren ein. Es handelt sich nicht um einen fairen Vergleich, da der Unterricht an den Schulen unter teilweise völlig unterschiedlichen Voraussetzungen stattfindet.

1.1.2 Vergleich mit Subgruppenmittelwerten

Die Berechnung von Leistungsmittelwerten für Subgruppen wird als marginale Adjustierung bezeichnet. Bei VERA werden die Landesergebnisse den Schulen häu-

1 International werden die Adjustierungsverfahren des fairen Vergleichs als *value-added models* bezeichnet (Leckie & Goldstein, 2017).

fig getrennt nach Schulart und Geschlecht rückgemeldet. So kann beispielsweise eine Schulleitung das Abschneiden der eigenen Schule mit der Gesamtleistung aller Gymnasien im Land vergleichen.

1.1.3 Vergleich mit ähnlichen (existierenden) Lerngruppen

Eine weitere Strategie ist der Vergleich mit ähnlichen Klassen bzw. Schulen. Die Bestimmung der Ähnlichkeit kann dabei unterschiedlich erfolgen. Im ersten Schritt wird in den meisten Fällen ein Index gebildet, der die relevanten Charakteristika der Klassen bzw. Schulen abbildet. In solch einen Index können Schulstandortinformationen, Belastungsindikatoren und weitere Kontextmerkmale einfließen. Die Indexbildung erfolgt zumeist, indem die gemeinsame Varianz der Merkmale auf einem eindimensionalen Index abgebildet wird (zum Beispiel durch Faktoranalysen oder andere Linearkombinationen). Im zweiten Schritt werden basierend auf dem Index die Vergleichsgruppen gebildet. Dies geschieht entweder durch die Einteilung der nach dem Index sortierten Schulen in Quantile oder durch die Auswahl derjenigen z. B. fünf Schulen, die auf dem Index am nächsten an der jeweiligen Zielschule liegen. Die Größe der Vergleichsgruppe schwankt somit stark je nach der Wahl des konkreten Vorgehens. Da in den Index häufig sozioökonomische und soziokulturelle Merkmale wie Migrationshintergrund, Alltagssprache, Bücherbestand sowie der Anteil von Bedarfsgemeinschaften mit Kindern im Schuleinzugsgebiet, die Arbeitslosengeld II nach dem Sozialgesetzbuch II beziehen, einfließen, spricht man meist von einem Sozialindex. Ein Sozialindex kann noch andere, weitere Funktionen haben wie etwa die Ressourcenallokation (Weishaupt, 2016).

1.1.4 Vergleich mit Erwartungswerten

Die modellbasierte Adjustierung der Vergleichswerte umfasst aus statistischer Sicht auch die drei vorherigen Strategien. Diese lassen sich als Regression der Testleistung auf die Faktorvariable Gruppenzugehörigkeit darstellen. Nimmt man die Kontextvariablen jedoch nicht zur Indexbildung, sondern als einzelne Prädiktoren in ein Regressionsmodell zur Vorhersage der Testleistung mit auf, so handelt es sich bei den daraus resultierenden Erwartungswerten nicht mehr um die (mittlere) Leistung existierender Schulen. Es wird anhand linearer Zusammenhänge ein theoretischer Vergleichswert geschätzt, „wie er für die spezifisch vorliegenden Kontextbedingungen zu erwarten ist“ (Nachtigall & Kröhne, 2006, S. 68). Die modellbasierte Adjustierung hat den Vorteil, dass die fairen Vergleichswerte klar definierte Standardfehler haben, dass verschiedene Ebenen (Individuum, Klasse, Schule) im selben (Mehrebenen-)Modell spezifiziert werden können und dass durch die Regressionsgewichte der Kontextvariablen die Fairness im Sinne der erklärten Leistungsvarianz gut erklärt werden kann (Pham, Robitzsch, George & Freundberger, 2016).

1.2 Empirische Vergleiche von Adjustierungsstrategien

Bei VERA gibt es zwischen den Bundesländern teils deutliche Unterschiede bezüglich der Adjustierungsstrategien und bezüglich der dabei berücksichtigten Kontextvariablen (vgl. Fiege, 2016). Systematische Untersuchungen, bei denen die Güte der verschiedenen Adjustierungsstrategien miteinander verglichen wird, sind jedoch selten. Diese beschränken sich zudem jeweils auf den Vergleich zwischen einzelnen Berechnungsvarianten und bieten teilweise nur äußerst lückenhafte Informationen (z. B. Isaac & Hosenfeld, 2008).

Die umfangreichste Analyse modellbasierter Adjustierungsmodelle (Fiege, 2014) basiert auf Daten des Thüringer Projekts *Kompetenztest.de* ($N = 12\,708$ Schüler:innen). Für die Adjustierung wurden in allen Modellen die Kontextvariablen Geschlecht, Klassenwiederholung, Muttersprache, sonderpädagogischer Förderbedarf, Bücherbestand und Schulart verwendet. Systematisch variiert wurden die Modellspezifikation (mit und ohne Interaktionstermen) und die Verwendung von Vorwissen als Kontextvariable. Die komplexeren Modelle (Interaktionen bzw. mehr Kontextvariablen) erhöhten dabei die Fairness im Sinne der Leistungsvarianzaufklärung bei VERA 8 in Deutsch und Mathematik. Der Zusatznutzen von Interaktionstermen verschwand jedoch bei der Hinzunahme von Vortestleistungen aus der dritten und sechsten Klasse.

Im Rahmen der Berliner ELEMENT-Studie fand ein Vergleich von Adjustierungsverfahren für Deutsch und Mathematik in den Klassen 2, 4 und 6 statt ($k = 74$, 69 und 68 Schulen; Kuhl, Lenkeit, Pant & Wendt, 2009). Bei der modellbasierten Adjustierung mittels Mehrebenenregression erklärten die Kontextvariablen Zuzahlungsbefreiung und Herkunftssprache auf Schulebene am meisten Varianz. Eine alternative Adjustierung erfolgte mittels Kontextgruppen. Dafür wurden basierend auf diesen beiden Variablen drei Gruppen gebildet (Schule auf beiden Variablen unter dem Median, Schule auf beiden Variablen über dem Median, Mischgruppe). Die erklärte Varianz wurde hierfür jedoch nicht berichtet. Stattdessen wurde betrachtet, wie häufig „die beiden Adjustierungsverfahren [...] zur gleichen Klassifikation der Schulen nach erwartungskonformen bzw. erwartungswidrigen Leistungsniveaus kommen“ (Kuhl et al., 2009, S. 251). Insgesamt gab es bei gut einem Drittel der Schulen voneinander abweichende Klassifikationen. Die Adjustierungsstrategien liefern also durchaus voneinander abweichende faire Vergleichswerte. Welches Modell fairer (im Sinne der Varianzaufklärung) ist, bleibt jedoch offen.

Bei der Überprüfung der österreichischen Bildungsstandards wurden verschiedene Regressionsvarianten für die modellbasierte Adjustierung auf Schulebene miteinander verglichen (z. B. Englisch in der 8. Klasse mit $k = 1402$ Schulen; Pham & Robitzsch, 2014). Auf eine Mehrebenenmodellierung wurde verzichtet, da diese für kleine Schulen Schätzungen liefert, die in Richtung des Gesamtmittelwertes verzerrt und somit nicht fair sind (vgl. Pham, Robitzsch et al., 2016). Für die Schätzung der Erwartungswerte wurden standortbezogene Merkmale (z. B. Urbanisierungsgrad), schulbezogene Merkmale (z. B. Schulart) sowie aggregierte Schülermerkmale (z. B.

mittlerer sozioökonomischer Status) als Kontextvariablen verwendet (Pham & Robitzsch, 2014).

1.3 Fairness und Praktikabilität

Trotz der unterschiedlichen Untersuchungsdesigns und Analysemethoden zeichnet sich in den bisherigen Studien zum fairen Vergleich ab, dass die erklärte Leistungsvarianz (R^2) ein hilfreiches Kriterium zur Einschätzung der Fairness und somit ein geeignetes Gütemaß für den Vergleich von Adjustierungsstrategien ist (vgl. Fiege, 2016). Die erklärte Leistungsvarianz steigt mit jedem zusätzlichen Prädiktor im Regressionsmodell. Insofern sind zum Beispiel Modelle mit Interaktionstermen Modellen ohne Interaktionsterme automatisch überlegen. Wichtig ist also nicht, ob, sondern wie weit zusätzliche Prädiktoren die Varianzaufklärung verbessern (Fiege, 2014; Pham, Robitzsch et al., 2016). Entsprechend wird ausgehend von den bisherigen Studien zum fairen Vergleich ein R^2-Zuwachs von weniger als einem Prozentpunkt ungeachtet von Signifikanztests als nicht bedeutsam bewertet. Außerdem ist zu beachten, dass sehr komplexe Modelle beziehungsweise kleine Stichproben zu einer Überanpassung führen können (Fiege, 2014).

Eine zusätzliche Möglichkeit, Aussagen über die Güte der geschätzten fairen Vergleichswerte zu treffen, bietet der Standardfehler der adjustierten Werte.[2] Er sollte möglichst klein ausfallen. Denn wenn bestimmte Datenbereiche nur mit wenigen Beobachtungen besetzt sind, lassen sich die entsprechenden Erwartungswerte nur mit großem Standardfehler schätzen (Fiege, 2014, S. 144). Es gibt keine kriteriale Norm, ab wann Standardfehler zu groß für eine reliable Schätzung sind. Der Vergleich der Standardfehler von verschiedenen Adjustierungsstrategien ist jedoch möglich. Bei der Adjustierung mittels Mehrebenenregression fallen die Standardfehler für die einzelnen Schüler:innen in der Regel sehr hoch aus, während es auf Schulebene hinreichend zuverlässige Erwartungswerte gibt. Für kleine Schulen und Schulen mit unüblichen Verteilungen der Kontextvariablen fällt der Standardfehler gewöhnlich höher aus (Fiege, 2014; Pham & Robitzsch, 2014).

In der Praxis steht der Fairnesszuwachs durch eine Adjustierung stets der Praktikabilität der Adjustierungen gegenüber (Fiege et al., 2011). Zur Praktikabilität gehören die Aspekte Erhebungsökonomie, Robustheit und Verständlichkeit.

Die Erhebung von Kontextvariablen ist meist mit Zusatzaufwand für alle Beteiligten verbunden (vgl. Weishaupt, 2016). Der Fairnessgewinn durch die Hinzunahme von Variablen muss entsprechend abgewogen werden mit den Kosten für die Erfassung. Die Verwendung von bestehenden Daten aus der amtlichen Schulstatistik erspart den Schulen Zusatzbelastung. Die Daten sind jedoch häufig nicht in der gewünschten inhaltlichen, räumlichen und zeitlichen Auflösung verfügbar. Einzelne Informationen lassen sich mit geringem Zusatzaufwand im Rahmen der Leis-

2 Dabei wird implizit angenommen, dass die Daten eine Stichprobe aus einer Grundgesamtheit (z. B. aller möglichen VERA-Durchgänge) sind.

tungstestung von den Schüler:innen sowie von den Lehrkräften erfragen (Kuhl et al., 2009). Auf diese Weise werden häufig Merkmale wie Alltagssprache, Geschlecht sowie der Bücherbestand als Indikator für den sozioökonomischen Status erhoben. Die Einschätzungen der Befragten sind zwar auf der Individualebene nur begrenzt zuverlässig, ergeben aggregiert auf Klassen- bzw. Schulebene jedoch ein realistisches Bild der Zusammensetzung der Schülerschaft (Isaac, 2008). Gezielte Zusatzbefragungen sind hingegen mit hohem Aufwand verbunden und werden in der Praxis nur eingesetzt, wenn die erhobenen Merkmale noch für andere Zwecke als den fairen Vergleich verwendet werden, etwa zur Ressourcenverteilung (vgl. Schulte, Hartig & Pietsch, 2016).

Die Kontextvariablen, die bei der Adjustierung verwendet werden, sollten robust gegenüber Manipulationen und anderen systematischen Verzerrungen sein. Bei den Angaben zum fairen Vergleich wie auch bei der Eingabe der eigentlichen Testergebnisse besteht die Gefahr, dass Daten beschönigt werden. Da das Abschneiden in VERA jedoch nicht mit Sanktionen verbunden ist, gibt es nur in Einzelfällen Falschangaben (vgl. Leutner, Fleischer, Spoden & Wirth, 2008; Spoden, Fleischer & Leutner, 2014). Zudem kann der faire Vergleich auch der Schulaufsicht bei der Einordnung der Leistungen und bei Gesprächen mit Schulleitungen helfen (Kemethofer & Wiesner, 2019, S. 236). Die meisten Adjustierungen für faire Vergleiche werden jährlich neu berechnet, wobei die Unterschiede innerhalb eines Durchgangs (z. B. zwischen Bundesländern) oft größer ausfallen als zwischen verschiedenen Durchführungsjahren (vgl. Pham, Freunberger, Robitzsch, Itzlinger-Bruneforth & Bruneforth, 2016; Pham & Robitzsch, 2014).

Zuletzt muss der faire Vergleich für die Zielpersonen verständlich sein. Nur wenn Lehrkräfte, Schulleitungen und Schulaufsicht die adjustierten Werte inhaltlich interpretieren und korrekt einordnen können, bietet der faire Vergleich einen Mehrwert und hat das Potenzial, zur Unterrichts- und Schulentwicklung beizutragen. Es zeigt sich zwar eine höhere Zufriedenheit und Akzeptanz von VERA in Bundesländern mit fairem Vergleich in der Ergebnisrückmeldung (Maier, Bohl, Kleinknecht & Metz, 2011), allerdings finden zumindest Grundschullehrkräfte den adjustierten Vergleichswert weniger nützlich als die Landes-, Schul- und Klassenwerte (Groß Ophoff, Koch & Hosenfeld, 2019).

1.4 Das Nutzen-Aufwand-Verhältnis

Obgleich bei VERA verschiedene Adjustierungsstrategien seit Jahren in vielen Bundesländern eingesetzt werden, gibt es kaum empirische Gegenüberstellungen hinsichtlich der erzielten Fairness und der Praktikabilität. Einheitliche Beurteilungsstandards für die Aspekte der Praktikabilität fehlen. Ebenso lässt sich das Nutzen-Aufwand-Verhältnis bislang nicht quantitativ beurteilen und abwägen. Bei der Wahl der verwendeten Kontextvariablen werden theoretische und pragmatische Überlegungen abgewogen (vgl. Kuhl et al., 2009). Aus theoretischer Sicht sind insbesondere soziales, ökonomisches und kulturelles Kapital sowie Migrationshinter-

grund mit sprachlichen und kulturellen Aspekten relevant (Schulte et al., 2016). Für einen möglichst fairen Vergleich sollte als Kontextvariable neben soziokulturellen Merkmalen möglichst fachspezifisches Vorwissen verwendet werden (Fiege, 2014). Die verwendeten Kontextvariablen sollten nicht nur auf (subjektiven) Lehrkraftangaben beruhen, da der Adjustierung sonst wenig Vertrauen geschenkt wird (Koch, 2011, S. 247). Wenngleich es keinen „Goldstandard" für die Auswahl der Kontextvariablen gibt, korrelieren die in der Praxis eingesetzten Variablen miteinander (s. hohe Primärfaktorladungen bei Schulte et al., 2016). Im Rahmen einer Kosten-Nutzen-Abwägung beschränkt sich die Anzahl der Kontextvariablen im flächendeckenden Einsatz meist auf Merkmale, die ohne große Zusatzbelastung der Befragten erfassbar sind. Die Adjustierung wird dabei robuster, wenn Daten aus unterschiedlichen Quellen gemeinsam verwendet werden. Angaben von Schüler:innen sowie von Lehrkräften und Schulleitungen sind im Rahmen der Testungen und der Schulstatistik effizient erfassbar. Elternfragebögen, wie sie in kleineren Studien häufig eingesetzt werden, kommen bei landesweiten Testungen aufgrund der zusätzlichen Logistik sowie der datenschutzrechtlichen Herausforderungen jedoch nur selten zum Einsatz (Kuhl et al., 2009).

Während einige Bundesländer wie etwa Hamburg für den fairen Vergleich die mittlere Leistung existierender Schulen rückmeldet, werden beispielsweise in Thüringen modellbasierte Adjustierungen vorgenommen. Der bisher einzige empirische Vergleich zeigte, dass je nach Modellierung eine von drei Schulen deutlich unterschiedliche faire Vergleichswerte erhält (Kuhl et al., 2009). Ein direkter Vergleich der erreichten Fairness im Sinne der erklärten Leistungsvarianz steht jedoch noch aus. In der vorliegenden Studie werden nun erstmals die wesentlichen Adjustierungsstrategien anhand einer landesweiten Vollerhebung bezüglich ihrer Fairness miteinander verglichen.

1.5 Forschungsfragen

Wie gut eignen sich die in der Praxis verwendeten Adjustierungsstrategien für den fairen Vergleich bei Lernstandserhebungen? Für VERA 3 und VERA 8 in Baden-Württemberg werden die gängigen Adjustierungen durchgeführt und hinsichtlich der erreichten Fairness im Sinne der erklärten Leistungsvarianz sowie der Standardfehler miteinander verglichen.

Faire Vergleiche können auf Klassenebene stattfinden, sind bei modellbasierten Adjustierungen dort bei kleinen Lerngruppen zum Gesamtmittelwert hin verzerrt und mit großen Standardfehlern behaftet (Pham, Robitzsch et al., 2016). Aus diesem Grund wird der faire Vergleich nachfolgend auf der Ebene von Schulen untersucht. Betrachtet werden die Fächer Deutsch (Teilbereich Lesen) und Mathematik (Globalskala). Diese Skalen basieren jeweils auf den Bildungsstandards der Kultusministerkonferenz. Sie bilden zentrale Kompetenzen der Grundschule und der Sekundarstufe I ab.

Als Kontextvariablen für den fairen Vergleich werden Geschlecht, Alltagssprache, Migrationshintergrund, Bücherbestand und Vorwissen (nur VERA 8) berücksichtigt. Auch wenn Geschlechtsunterschiede bei Schulleistungen meist klein ausfallen (Fischer, Schult & Hell, 2013), wird die Quote häufig beim fairen Vergleich verwendet (Fiege, 2014). Die Alltagssprache der Schüler:innen ist relevant, da eine mangelnde Beherrschung der Unterrichtssprache zu geringeren Lernerfolgen führen kann (Kempert et al., 2016). Als ergänzender Aspekt wird der Anteil der Schüler:innen mit Migrationshintergrund berücksichtigt. Nach der Definition der Kultusministerkonferenz liegt ein Migrationshintergrund vor bei Schüler:innen mit ausländischer Staatsangehörigkeit, mit Geburtsort außerhalb Deutschlands oder mit überwiegend ausländischer Verkehrssprache in der Familie bzw. im häuslichen Umfeld. Trotz der inhaltlichen Überlappung ist Migrationshintergrund auch nach der Kontrolle für den Unterrichtssprachgebrauch systematisch mit Schulleistung assoziiert und somit für einen fairen Vergleich von Bedeutung (Schulte et al., 2016). Die *Bücherfrage* ist ein zeitstabiler Indikator des bildungsrelevanten soziokulturellen Kapitals in Hinblick auf den familiären Bildungshintergrund (Schwippert, 2019). Von diesem hängen die bildungsrelevanten Ressourcen ab, die in der Familie zur Verfügung stehen, weshalb er ein wichtiger Aspekt bei der Berechnung des fairen Vergleichswerts ist (Fiege, 2014). Bei VERA 8 wird zudem das Vorwissen als Kontextvariable verwendet. Dafür werden die Schulergebnisse aus einer Lernstandserhebung zu Beginn der 5. Klasse berücksichtigt, um die fairen Vergleichswerte auch für die Lernausgangslage der Schüler:innen zu adjustieren.

2. Methode

2.1 Stichprobe und Studiendesign

Die Forschungsfragen wurden anhand der Daten von VERA 3 und VERA 8 aus den Jahren 2019 und 2021 in Baden-Württemberg[3] untersucht. Die Vergleichsarbeiten VERA sind bundesweite schriftliche Leistungstests zur Ermittlung des Kompetenzstands von Schüler:innen im zweiten Schulhalbjahr der dritten bzw. achten Klasse bezüglich ausgewählter Kompetenzbereiche aus den Bildungsstandards u. a. in den Fächern Deutsch und Mathematik. Die Arbeiten werden vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) entwickelt. Die Durchführung in Baden-Württemberg obliegt dem Institut für Bildungsanalysen Baden-Württemberg (IBBW). Die Teilnahme war verpflichtend für alle öffentlichen Grundschulen (VERA 3) bzw. für alle öffentlichen allgemeinbildenden Schulen der Sekundarstufe I. Im Jahr 2019 nahmen $K_{\text{VERA3}} = 2330$ Grundschulen und $K_{\text{VERA8}} = 1261$ weiterführende Schulen verpflichtend teil; im Jahr 2021 waren es $K_{\text{VERA3}} = 2316$ Grundschulen und $K_{\text{VERA8}} = 1219$ weiterführende Schulen. Die Zahl der teilnehmenden Schüler:innen

3 Siehe https://ibbw-bw.de/Ergebnisse+Fremdevaluation+_+VERA. Details zum jeweiligen VERA-Durchgang sind hier abrufbar.

reichte von $n = 81\,491$ (in VERA 3 Deutsch 2021) bis $n = 84\,334$ (in VERA 8 Mathematik 2019). Da die Testungen 2021 pandemiebedingt erst im September des folgenden Schuljahrs (VERA 3 in Klasse 4 und VERA 8 in Klasse 9) durchgeführt wurden, liegt der Fokus der vorliegenden Analysen auf dem Jahr 2019. Die Daten von 2021 werden deshalb zur ergänzenden Validierung betrachtet.

2.2 Leistungsmessung und Kontextvariablen

Die Tests wurden während der regulären Schulzeit geschrieben. Die jeweilige Fachlehrkraft der Klasse korrigierte die Tests mithilfe standardisierter Auswertungsanleitungen. Richtig gelöste Items wurden mit 1 gewertet, falsch gelöste und nicht bearbeitete Items wurden mit 0 gewertet. Die Leistung der Schüler:innen wurde mit dem Rasch-Modell geschätzt (weighted likelihood estimation) und anschließend auf Schulebene gemittelt. Die Aufgabenschwierigkeiten wurden anhand von Normstichproben für die Bildungsstandards und von zeitlich vorgelagerten Pilotierungen ermittelt. Die Leistungen konnten so auf die Metrik der Bildungsstandards transformiert werden. Die Leistungen in der Normstichprobe hatten jeweils den Mittelwert 500 (in VERA 8 Mathematik 525) und die Standardabweichung 100.[4] Die Eigenschaften der Leistungstests sind in Tabelle 1 zusammengefasst.

Tabelle 1: Eigenschaften der Testhefte in Baden-Württemberg

Jahr	Testbereich	Testzeit	Itemanzahl Heft 1	Itemanzahl Heft 2
2019	VERA 3 Deutsch (Lesen)	40 Minuten	22	–
	VERA 3 Mathematik (Globalskala)	60 Minuten	34	–
	VERA 8 Deutsch (Lesen)	40 Minuten	41	38
	VERA 8 Mathematik (Globalskala)	80 Minuten	48	52
2021	VERA 3 Deutsch (Lesen)	40 Minuten	21	–
	VERA 3 Mathematik (Globalskala)	40 Minuten	25	–
	VERA 8 Deutsch (Lesen)	35/40 Minuten	21	29
	VERA 8 Mathematik (Globalskala)	60 Minuten	35	29

Anmerkungen. Bei VERA 3 Mathematik 2019 wurden die Leitideen *Daten, Häufigkeit und Wahrscheinlichkeit* sowie *Raum und Form* getestet und gemeinsam skaliert. In VERA 8 wurden zwei Testheftversionen mit überlappenden Aufgaben eingesetzt: Heft 2 an Gymnasien, Heft 1 in nichtgymnasialen Schularten.

Die Angaben zum Geschlecht (männlich/weiblich) und zur Alltagssprache (Deutsch/nicht Deutsch) der Schüler:innen stammen von der jeweiligen Fachlehrkraft. Der Anteil der Schüler:innen in der 3. bzw. 8. Klassenstufe mit Migrationshintergrund stammt aus der Schulstatistik.

4 Die Skalierung basiert auf den Kompetenzstufenmodellen der jeweiligen Bildungsstandards der Primarstufe (VERA 3) bzw. des Mittleren Schulabschlusses (VERA 8) – Details siehe https://www.iqb.hu-berlin.de/bista/ksm.

Der Bücherbestand wurde analog zum IQB-Bildungstrend bei VERA 3 mit einer fünfstufigen und bei VERA 8 mit einer sechsstufigen Antwortskala direkt von den Schüler:innen zu Beginn des Deutsch-Testhefts erfragt. Bei VERA 3 wurden die Antwortoptionen mit Abbildungen von Bücherregalen illustriert.

Als Indikator für das Vorwissen auf Schulebene wird bei VERA 8 die mittlere Leistung verwendet, die die untersuchten Kohorten zuvor jeweils im Lernstand 5 erreichten (Lernstand 5 2015 für VERA 8 2019 sowie Lernstand 5 2017 für VERA 8 2021). Lernstand 5 testet Basiskompetenzen in Deutsch und Mathematik zu Beginn der 5. Klasse an allen öffentlichen weiterführenden allgemeinbildenden Schulen in Baden-Württemberg (vgl. Schult, Mahler, Fauth & Lindner, 2022).

Der Lesetest in Deutsch bestand jeweils aus vier Texten und 32 bzw. 38 dazugehörigen Items. Die Bearbeitungszeit betrug 50 Minuten. Die Aufgaben erfassten Leseprozesse wie die Identifikation einfacher Informationen, die Verknüpfung von Informationen aus dem Text, die Formulierung von Begründungen sowie komplexen Schlussfolgerungen. Der Mathematiktest umfasste drei jeweils 20-minütige Aufgabenblöcke. Die schriftlichen Rechenverfahren bestanden aus 12 Items zur Subtraktion, Multiplikation und Division. Das Operationsverständnis (14 Items) bezog sich auf Kompetenzen wie Operationen in Alltagssituationen anwenden und teils mehrschrittige Operationen verstehen und flexibel anwenden. Das Zahlverständnis (14 Items) umfasste den Umgang mit Stellenwerten, Vorstellungen zu Zahlengrößen und das Verständnis von Zahlen in problemhaltigen Situationen. Als Leistung wurde der (mittlere) Anteil der gelösten Aufgaben im Lesen (Deutsch) bzw. im Gesamttest (Mathematik) verwendet.

2.3 Statistische Modelle der Adjustierung

Als Referenzmodell für die Adjustierungen diente die Rückmeldung des Landesmittelwerts (VERA 3, Strategie I bei Fiege, 2014, S. 72) bzw. des schulartspezifischen Landesmittelwerts (VERA 8). Basierend auf den Kontextvariablen Geschlecht, Alltagssprache, Migrationshintergrund und Bücherbestand wurde eine Faktoranalyse durchgeführt, deren erster Faktor als Index für den soziokulturellen Status verwendet wurde. In Modell 1 wurden die Schulen nach diesem Index sortiert und in 5-%-Perzentile eingruppiert (Strategie IIIc bei Fiege, 2014, S. 72) – bei VERA 8 getrennt nach Schulart. Entsprechend betrug die durchschnittliche Größe der einzelnen Kontextgruppen 117 Schulen bei Grundschulen (VERA 3), 15 bei Werkreal-/Hauptschulen, 14 bei Gemeinschaftsschulen, 20 bei Realschulen und 19 bei Gymnasien (VERA 8). In Modell 2 wird die gleiche Strategie angewendet mit der Einteilung in 10-%-Perzentile.[5] Als adjustierten Wert bekam jede Schule den Mittelwert aller Schulen in ihrer Perzentil-Gruppe. In Modell 3 wurden die Schulen nach

5 In Bundesländern, die diese Strategie nutzen, liegt die Anzahl der Kontextgruppen im unteren einstelligen Bereich. Eine vertiefte Auswertung der vorliegenden Daten zeigte jedoch, dass für Baden-Württemberg die Fairness dieser Adjustierungsstrategie mit der Anzahl der Kontextgruppen tendenziell zunimmt und eine Einteilung in 5-%-Perzentil-

dem Index sortiert und erhielten als adjustierten Wert die mittlere Leistung der 10 Schulen, die jeweils auf dem Index direkt oberhalb und unterhalb der Zielschule lagen. Der Vergleichswert basiert damit insgesamt auf 20 Schulen (Strategie IIIb bei Fiege, 2014, S. 72). In den Stadtstaaten Berlin und Hamburg basiert der Vergleichswert auf 8 bzw. 6 Schulen. Im Flächenland Baden-Württemberg sind Adjustierungen mit dieser kleinen Anzahl an Vergleichsschulen jedoch nicht so fair wie für größere Vergleichsgruppen aus 20 Schulen (Schult, 2020).

In Modell 4 wurde mit einfachen linearen Regressionen die Schulleistung durch den Index als unabhängige Variable vorhergesagt. In Modell 5 wurden lineare Regressionen zur Vorhersage der Schulleistung durchgeführt mit den Prädiktoren Schulart, Geschlecht, Alltagssprache, Migrationshintergrund und Bücherbestand. Die Erwartungswerte aus den Regressionsmodellen wurden als adjustierte Werte verwendet (Strategie IVa bei Fiege, 2014, S. 72).

Für VERA 8 wurden noch weitere Modelle betrachtet: Analog zum Vorgehen von Fiege (2014) wurden lineare Regressionsmodelle geschätzt, bei denen die Prädiktoren mit der Schulart interagieren (Modell 6 und Modell 9). In Modell 7 wurde als einziger Prädiktor das Vorwissen in eine einfache Regression aufgenommen (value-added model zur Berücksichtigung von Vorkenntnissen; Leckie & Goldstein, 2017). In Modell 8 wurde in den linearen Regressionen von Modell 5 zusätzlich der Prädiktor Vorwissen aufgenommen (Strategie IVb bei Fiege, 2014, S. 72).

2.4 Analyseplan

Die Analysen fanden auf Schulebene statt. Alle adjustierten Werte ($\hat{y}_i$) sowie die dazugehörigen Standardfehler wurden durch Regressionsmodelle geschätzt, die jeweils nach Anzahl der teilnehmenden Schüler:innen (n_i) gewichtet wurden. Als Maß für die Fairness wurde der Determinationskoeffizient R^2 verwendet gemäß der Formel $R^2 = (\sigma - \sigma^*)/\sigma$ mit der Gesamtleistungsvarianz σ, dem gewichteten Mittel der quadrierten Residuen $\sigma^* = n_i(y_i - \hat{y}_i)^2 / \sum n_i$. Da die Adjustierungsmodelle nicht alle ineinander genestet sind und da es sich um die Analyse einer Vollerhebung handelt, fand kein inferenzstatistischer Modellvergleich statt. Stattdessen wurden ausgehend von den bisherigen Arbeiten R^2-Unterschiede ab einem Prozentpunkt und beim mittleren Standardfehler ab einem Punkt auf der Bildungsstandardskala als bedeutsam interpretiert.

Fehlende Schulwerte beim Vorwissen wurden mittels multipler Imputation ersetzt mit den anderen Modellvariablen als Hintergrundinformation und jeweils fünf imputierten Datensätzen ($k < 133$)[6]. Fehlende Schulwerte beim Bücherbestand ($k < 15$) wurden durch den (schulartspezifischen) Landesmittelwert ersetzt. Fehlen-

gruppen entsprechend fairere Vergleichswerte liefert als beispielsweise 25-%-Perzentilgruppen (Schult, 2020).

6 Da bei der Zuordnung der Lernstand-5-Ergebnisse die Schulart bei Schulverbünden nachträglich nur teilweise möglich war, ist die Zahl der fehlenden Werte leider recht hoch.

de Werte beim Migrationshintergrund auf Klassenstufenebene ($k < 200$) wurden durch den Schulwert ersetzt. Die Analysen wurden in R 4.0.4 (R Core Team, 2021) mit den Paketen *ggplot2* (Wickham, 2016) und *mice* (van Buuren & Groothuis-Oudshoorn, 2011) umgesetzt.

3. Ergebnisse

3.1 VERA 3

Der Determinationskoeffizient R^2 als Gütemaß der Fairness war bei VERA 3 am höchsten für die Regression der Testleistung auf die Prädiktoren Geschlecht, Alltagssprache, Migrationshintergrund und Bücherbestand (Modell 5). In Deutsch wie auch in Mathematik war die erklärte Varianz für Modell 5 im Jahr 2019 mit 46% bzw. 33% jeweils mindestens 7 Prozentpunkte (und somit bedeutsam) höher als für die indexbasierten Adjustierungsstrategien (Modell 1 bis Modell 4; vgl. Tabelle 2). Im Jahr 2021 zeigt sich das gleiche Muster mit 45% bzw. 38% erklärter Varianz für Modell 5.

Tabelle 2: Vergleich der Adjustierungsstrategien hinsichtlich der Fairness (R^2) und des mittleren Standardmessfehlers ($\overline{SE}$) bei VERA 3

		2019 ($k = 2330$)				2021 ($k = 2316$)			
		Deutsch		Mathematik		Deutsch		Mathematik	
Modell		R^2	$\overline{SE}$	R^2	$\overline{SE}$	R^2	$\overline{SE}$	R^2	$\overline{SE}$
0	Landesmittelwert	0	1	0	0.9	0	1	0	1
1	Index-5-%-Perzentilgruppen	.361	3.6	.253	3.6	.346	3.7	.287	3.6
2	Index-10-%-Perzentilgruppen	.331	2.6	.227	2.6	.321	2.7	.264	2.6
3	Indexnachbarschulen	.337	8.5	.215	8.6	.312	8.7	.245	8.6
4	Einfache Regression mit Index	.367	1.1	.257	1.1	.348	1.1	.291	1.1
5	Regression	.464	1.6	.326	1.6	.452	1.6	.383	1.6

Hinsichtlich der Standardfehler der adjustierten Werte hat Modell 4, die einfache Regression mit dem Index als Prädiktor, in beiden Fächern mit durchschnittlich jeweils 1.1 Punkten auf der Bildungsstandardskala den niedrigsten und somit günstigsten Wert (abgesehen freilich vom nicht adjustierten Landesmittelwert; vgl. Tabelle 2). Während das sehr faire Modell 5 mit 1.6 etwas größere mittlere Standardfehler aufweist, war er bei den Modellen 1 bis 3 hingegen mit über 2.5 deutlich höher. Besonders hoch fiel er bei Modell 3 aus, da immer nur eine geringe Zahl an Schulen bei dieser Adjustierung berücksichtigt wurde. Die Verteilung der Standardfehler der einzelnen adjustierten Schulwerte wird in Abbildung 1 illustriert. Bei Modell 5 gibt es ein paar wenige Ausreißer nach oben, die jedoch alle noch im einstelligen Punktebereich liegen. Die Standardfehler korrelieren bei Modell 5 negativ mit der Schülerzahl (2019: in beiden Fächern jeweils $r = -.29$; 2021: jeweils $r = -.28$).

Entsprechend handelt es sich bei den Ausreißern um Schulen mit geringer Schülerzahl. Das Streudiagramm in Abbildung 2 zeigt allerdings, dass Modell 5 auch bei zahlreichen kleinen Schulen nur kleine Standardfehler aufweist.

Abbildung 1: Boxplots mit den Verteilungen der Standardfehler der adjustierten Schulwerte für VERA 3 (2019)

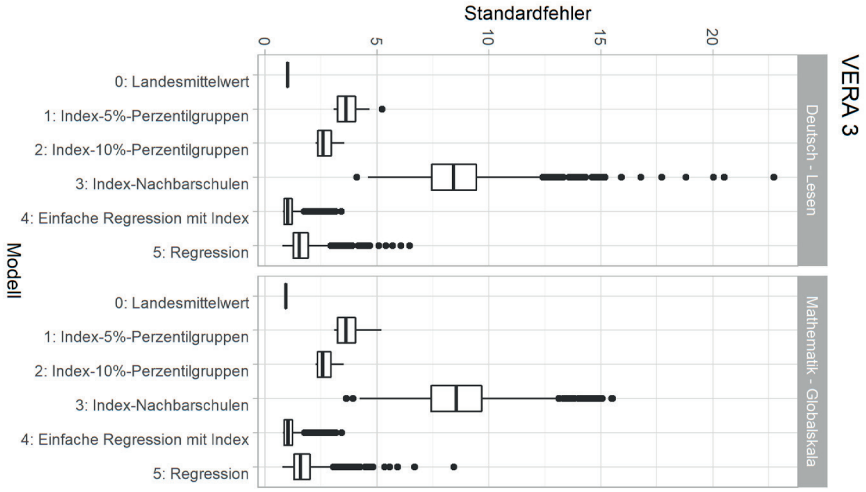
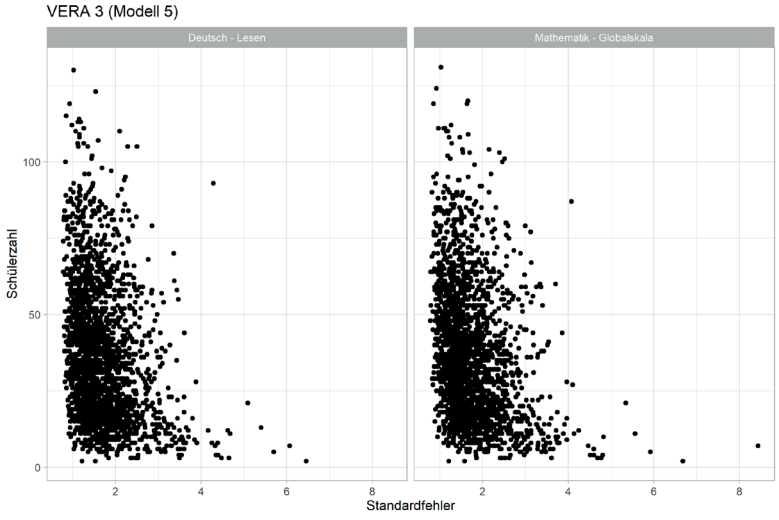


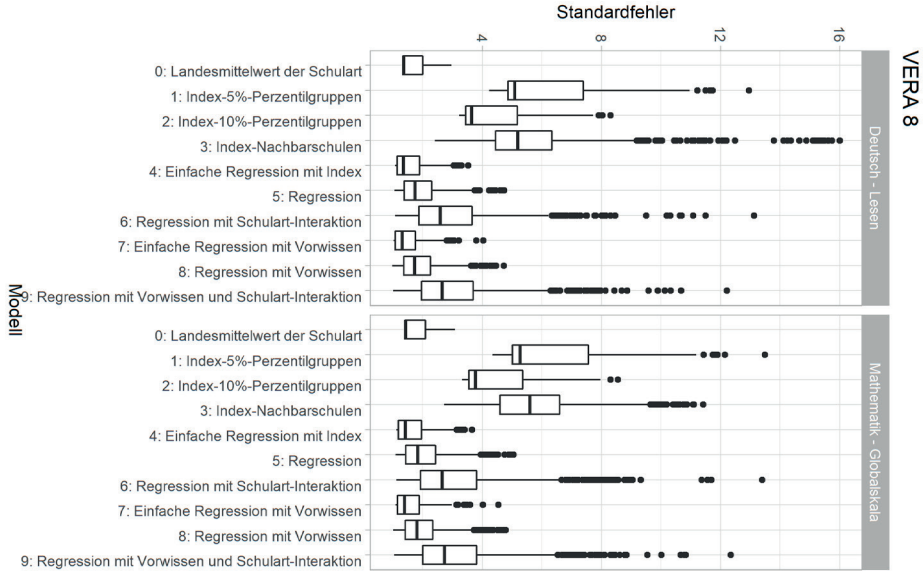
Abbildung 2: Streudiagramm der Standardfehler von Modell 5 und der Anzahl der jeweils teilnehmenden Schüler:innen pro Schule für VERA 3 (2019)



3.2 VERA 8

Bei VERA 8 ergab sich das größte R^2 für die Regression der Testleistung auf die Kontextvariablen Geschlecht, Alltagssprache, Migrationshintergrund, Bücherbestand, Vorwissen sowie alle Schulartinteraktionen dieser Prädiktoren (Modell 9, vgl. Tabelle 3). In Deutsch wie auch in Mathematik war die erklärte Varianz für das entsprechende Modell ohne Interaktionsterme (Modell 8) im Jahr 2019 mit 93% bzw. 92% allerdings nur um maximal 0.2 Prozentpunkte kleiner. Durch Weglassen des Vorwissens als Prädiktor (Modell 5) sinken die R^2-Werte jedoch um mehr als 1 Prozentpunkt. Der Kontextgruppenvergleich mit 5-%-Perzentilgruppen (Modell 1) erreicht eine vergleichbare Fairness wie Modell 5, während der Vergleich mit Indexnachbarschulen (Modell 3) sowie das reine value-added model mit ausschließlich Vorwissen als Prädiktor (Modell 7) etwas schlechter abschneiden als Modell 5 und Modell 8 (vgl. Tabelle 3). Für 2021 zeigt sich hinsichtlich der verschiedenen Adjustierungsstrategien das gleiche Befundmuster wie für 2019.

Abbildung 3: Boxplots mit den Verteilungen der Standardfehler der adjustierten Schulwerte für VERA 8 (2019)



Hinsichtlich der Standardfehler der adjustierten Werte hat Modell 7 mit durchschnittlich 1.3 Punkten die niedrigsten und somit günstigsten Werte, gefolgt von Modell 4 mit maximal 1.5 Punkten in beiden Jahren (vgl. Tabelle 3). Mit bis zu 1.8 Punkten fiel der mittlere Standardfehler für die Modelle 5 und 8 ebenfalls niedrig aus. Die Hinzunahme von Interaktionstermen erhöhte die Standardfehler im Schnitt um knapp 1 Punkt. Bei den indexbasierten Adjustierungen (Modell 1 bis 3) war der mittlere Standardfehler jeweils um mindestens 2 Punkte auf der Bildungsstandardskala größer als bei Modell 5 und Modell 8. Die Verteilung der Standard-

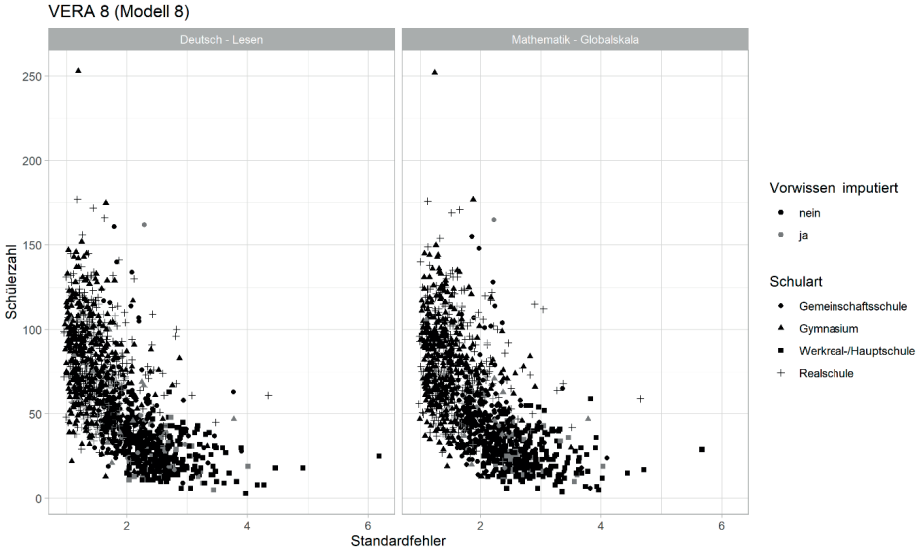
fehler der adjustierten Schulwerte wird in Abbildung 3 illustriert. Bei Modell 8 gibt es 2019 jeweils einen Ausreißer mit einem Standardfehler von (knapp) über 5 Punkten. Die Standardfehler aus Modell 8 korrelieren negativ mit der Schülerzahl (2019: $r_{\text{Deutsch}} = -.64$, $r_{\text{Mathematik}} = -.62$; 2021; beide $r = -.61$). Entsprechend handelt es sich bei dem Ausreißer um eine eher kleine Werkreal-/Hauptschule, die zudem einen der niedrigsten Werte beim Bücherbestand aufweist. Das Streudiagramm in Abbildung 2 illustriert den Zusammenhang von Standardfehlern und Schülerzahlen und zeigt zudem, dass die Imputation beim Vorwissen nicht mit erhöhten Standardfehlern zusammenhängt.

Tabelle 3: Vergleich der Adjustierungsstrategien hinsichtlich der Fairness (R^2) und des mittleren Standardmessfehlers ($\overline{SE}$) bei VERA 8

		2019 ($k = 1348^a$)				2021 ($k = 1301^a$)			
		Deutsch		Mathematik		Deutsch		Mathematik	
Modell		R^2	$\overline{SE}$	R^2	$\overline{SE}$	R^2	$\overline{SE}$	R^2	$\overline{SE}$
0	Landesmittelwert	0	2.2	0	2.1	0	1.9	0	2.3
0	Landesmittelwert der Schulart	.869	1.5	.866	1.5	.815	1.6	.861	1.7
1	Index-5-%-Perzentilgruppen	.917	5.5	.908	5.7	.887	5.7	.918	5.9
2	Index-10-%-Perzentilgruppen	.913	3.9	.903	4.1	.881	4.1	.914	4.2
3	Indexnachbarschulen	.908	5.1	.898	5.3	.873	5.2	.907	5.4
4	Einfache Regression mit Index	.912	1.4	.902	1.4	.879	1.4	.911	1.5
5	Regression	.920	1.6	.909	1.7	.893	1.7	.919	1.8
6	Regression mit Schulart-Interaktion	.922	2.5	.911	2.6	.894	2.5	.922	2.6
7	Einfache Regression mit Vorwissen	.920	1.3	.915	1.3	.894	1.3	.917	1.4
8	Regression mit Vorwissen	.932	1.6	.924	1.7	.908	1.7	.932	1.7
9	Regression mit Vorwissen und Schulart-Interaktion	.933	2.5	.926	2.6	.910	2.6	.934	2.7

Anmerkung. [a] Unter den analysierten Schulen bei VERA 8 waren Schulverbünde, aus denen mehrere Schularten separat in die Analysen einflossen.

Abbildung 4: Streudiagramm der Standardfehler von Modell 8 und der Anzahl der jeweils teilnehmenden Schüler:innen pro Schule für VERA 8 (2019)



4. Diskussion

Die Adjustierung im Rahmen des fairen Vergleichs liefert den Schulen eine soziale Bezugsnorm, die im Gegensatz zum Vergleich mit dem Landesmittelwert die Rahmensituation der Schule berücksichtigt. In der vorliegenden Studie wurden Adjustierungsstrategien, die bereits in verschiedenen Bundesländern eingesetzt werden, anhand von Daten aus Baden-Württemberg hinsichtlich ihrer Fairness verglichen. Bei VERA 3 schnitten modellbasierte Erwartungswerte jeweils besser ab als indexbasierte Adjustierungen. Bei VERA 8 fiel dieser Vorsprung aufgrund der Berücksichtigung der Schulart in allen Modellen deutlich geringer aus. Hinsichtlich der mittleren Standardfehler der adjustierten Werte war die Adjustierung durch modellbasierte Erwartungswerte (Regression mit den Kontextvariablen als Einzelprädiktoren) der indexbasierten Adjustierung jeweils in beiden Klassenstufen und beiden Fächern überlegen. Im Einklang mit VERA-8-Befunden aus Thüringen (Fiege, 2014) verbesserte die Verwendung von Vorwissen als zusätzlichem Prädiktor neben soziokulturellen Kontextvariablen die Fairness der Adjustierung, während die Modellierung von Schulartinteraktionen keinen Mehrwert bot.

Die klaren Befunde der vorliegenden Studie vervollständigen die Forschung zum fairen Vergleich, indem die verschiedenen Adjustierungsstrategien gemeinsam betrachtet werden. War bislang nur bekannt, dass die adjustierten Werte in Abhängigkeit der gewählten Strategie voneinander abweichen (Kuhl et al., 2009), so zeigt sich nun, dass die Fairness für indexbasierte Verfahren systematisch geringer ausfällt als für modellbasierte Erwartungswerte. Das Befundmuster ist über die beiden untersuchten Jahre, Klassenstufen und Fächer hinweg stabil und deckt sich hin-

sichtlich der Relevanz des Vorwissens mit zurückliegenden Untersuchungen (Fiege, 2014). Dies deutet die Verallgemeinerbarkeit der vorliegenden Ergebnisse an.

Bei der Modellierung der verschiedenen Adjustierungsstrategien als Regressionsmodelle deutet sich bereits an, dass es bei indexbasierten Verfahren zu einem höheren Informationsverlust durch die Indexbildung und zu größeren Standardfehlern beim Vergleich mit nur wenigen Vergleichsschulen kommt. Der geringeren Fairness der indexbasierten Verfahren steht möglicherweise eine bessere Verständlichkeit und somit eine höhere Praktikabilität gegenüber (vgl. z. B. die Darstellung bei Emmrich, Ernst, Harych & Wesselhöft, 2012). Allerdings ist die tatsächliche Indexberechnung auch nicht trivial (vgl. Schulte et al., 2016) und die Annahme der Eindimensionalität ergibt ein vereinfachtes Bild der sozialen Hintergründe. Die Indexhintergründe werden in den Erläuterungspapieren für Lehrkräfte nicht dargestellt. Generell scheint es sinnvoll, in Hinweispapieren für Lehrkräfte und Schulleitungen auf die inhaltliche Interpretation zu fokussieren (z. B. Isaac, 2008) und die Dokumentation der Berechnungen in technische Berichte auszulagern (z. B. Pham & Robitzsch, 2014).

4.1 Praxisrelevanz der Befunde

Die Landesmittelwerte sind oftmals schon ein fester Bestandteil der Rückmeldungen. Entsprechend geht es in erster Linie darum, einen fairen Vergleichswert als zusätzliche Information bereitzustellen. Da Schulpersonal mangels Zeit und Erfahrung kaum in der Lage ist, die Fairness verschiedener Adjustierungsverfahren abzuwägen, obliegt es den entsprechenden Institutionen, welche die Lernstandserhebungen durchführen, ein geeignetes Verfahren auszuwählen und zu erläutern.

Bei der Bildung von Kontextgruppen (Modell 1) kann die Gruppenzuweisung gerade am unteren Indexende als stigmatisierend empfunden werden. Jede Schule innerhalb einer Gruppe erhält den gleichen adjustierten Vergleichswert unabhängig davon, wie weit sie auf dem Index nach oben bzw. unten vom Gruppenmittelwert abweicht. Die tatsächliche Indexposition ist dabei für die Schule in der Regel nicht transparent.

Beim Vergleich mit einigen wenigen Schulen mit ähnlichen Indexwerten zeigten sich empirisch die niedrigsten Determinationskoeffizienten sowie die größten Standardfehler. Da eine datenbasierte Unterrichts- und Schulentwicklung eine möglichst informative Datenbasis braucht, scheint diese Adjustierungsstrategie entsprechend ungünstig, da die rückgemeldeten Vergleichswerte nur bedingt den sozialen Kontext und die Lernsituation an der Schule reflektieren.

Bei modellbasierten Erwartungswerten aus linearen Regressionen fielen die Gütemaße der Fairness in beiden untersuchten Klassenstufen in beiden Fächern hingegen am günstigsten aus. Im Rahmen einer Untersuchung an 16 hessischen Gymnasien bewerteten 9 von 19 befragten Lehrkräften den fairen Vergleich in dieser Form als interessantesten Teil der Ergebnisrückmeldung (Korngiebel, 2014, S. 191), was die Praxistauglichkeit dieser Adjustierungsstrategie unterstreicht. Eine Doku-

mentation der Adjustierung, wie sie beispielsweise in Österreich stattfindet (Pham & Robitzsch, 2014), sollte dabei Nachfragen zur genauen Berechnung und Transparenz beantworten.

4.2 Limitationen

Die vorliegende Untersuchung basiert auf Daten aus einem Flächenbundesland, während bisherige Forschungsarbeiten teilweise aus Stadtstaaten stammen (Isaac & Hosenfeld, 2008; Kuhl et al., 2009). Aufgrund kleinerer Stichproben und geringerer Varianz bei den Kontextvariablen mit Urbanisierungsbezug könnten sich die Befundmuster in Bundesländern wie Berlin, Bremen und Hamburg möglicherweise verschieben. Aufgrund der Gliederung des baden-württembergischen Schulsystems mit vier allgemeinbildenden Schularten in der Sekundarstufe I fallen die R^2-Werte bei VERA 8 jedenfalls höher aus als in Studien zum fairen Vergleich aus anderen Bundesländern. Die Auswahl der Kontextvariablen hängt ebenfalls von bundeslandspezifischen Begebenheiten ab. Beispielsweise wurde in der Berliner ELEMENT-Studie die Lernmittelzuzahlungsbefreiung berücksichtigt, die es aufgrund der allgemeinen Lernmittelfreiheit in Baden-Württemberg nicht gibt. Aufgrund der Erhebungsökonomie sowie der Korrelationen verschiedener sozioökonomischer und soziokultureller Kontextvariablen scheinen die untersuchten Kontextvariablen für Baden-Württemberg jedoch inhaltlich angemessen und mit verhältnismäßigem Erhebungsaufwand verbunden zu sein.

Bei jeder Adjustierungsstrategie gibt es zahlreiche Stellschrauben. Die hier gewählten Modelle orientieren sich stark an der Praxis, sodass eine umfassende Modellierung kausaler Effekte nicht möglich ist und auch nicht angestrebt wird. Der faire Vergleich bleibt somit ein *fairerer* Vergleich (Fiege, 2016). Beispielsweise bildet die Eindimensionalität des verwendeten Index die soziale Zusammensetzung vermutlich nur unvollständig ab.

Die Ergebnisse auf Schulebene sagen nur bedingt etwas über Klassen- und Individualeffekte aus. Bei der Modellierung auf Klassen- und Schülerebene können systematische Verzerrungen auftreten. Bei kleinen Klassen tendieren die adjustierten Erwartungswerte zum Gesamtmittelwert (Pham & Robitzsch, 2016) und Messfehler bei der Vortestleistung verzerren die Prädiktion von Schülerleistungen (Televantou et al., 2015). Die Adjustierungen auf Schulebene sind dagegen robuster gegenüber systematischen Verzerrungen, sodass die adjustierten Werte aus der vorliegenden Studie für die Ergebnisrückmeldung in der Praxis geeignet sind (vgl. auch die Schulrückmeldung in Österreich; Pham, Robitzsch et al., 2016).

5. Fazit

In seiner Meta-Meta-Analyse zu Determinanten des Lernerfolgs trennte Hattie (2009) zwischen Faktoren, auf die Lehrkräfte und im weiteren Sinne Entschei-

dungspersonen im Schulbereich direkten Einfluss nehmen können, und Faktoren, die jenseits des schulischen Gestaltungsspielraums liegen (z. B. Alltagsbildung im Lernort Familie). Beim fairen Vergleich wird versucht, diese außerschulischen Einflussfaktoren zu berücksichtigen, sodass Lehrkräfte und Schulleitungen neben dem Landeswert zusätzlich einen adjustierten Erwartungswert rückgemeldet bekommen, der angibt, wie Klassen bzw. Schulen mit vergleichbaren Ausgangsbedingungen abgeschnitten haben. Der vorliegende Vergleich verschiedener Adjustierungsstrategien zeigt, dass die modellbasierte Regression mit soziokulturellen Kontextvariablen (bei VERA 8 zusätzlich mit Schulart und Vorwissen) im Vergleich zu anderen Adjustierungsstrategien das höchste Maß an Fairness und Schätzgenauigkeit bietet.

Ein fairer Vergleich alleine macht zwar noch kein faires Bildungssystem. Mithilfe des fairen Vergleichs kann die VERA-basierte Unterrichts- und Schulentwicklung aber zumindest soziale Heterogenität und Leistungsheterogenität aufzeigen und Impulse liefern für einen professionellen Umgang mit Vielfalt.

Literatur

Dumont, H., Neumann, M., Nagy, G., Becker, M., Rose, N. & Trautwein, U. (2013). Einfluss der Klassenkomposition auf die Leistungsentwicklung in Haupt- und Realschulen in Baden-Württemberg. *Psychologie in Erziehung und Unterricht, 60*(3), 198–213. https://doi.org/10.2378/peu2013.art16d

Emmrich, R., Ernst, C.-M., Harych, P. & Wesselhöft, K. (2012). *Vergleichsarbeiten der Jahrgangsstufe 8 in Berlin als Beitrag zur Schul- und Unterrichtsentwicklung* (2. Aufl.). Institut für Schulqualität der Länder Berlin und Brandenburg. Verfügbar unter https://www.isq-bb.de/wordpress/wp-content/uploads/2016/05/Broschure_BLN_VERA_8_2011.pdf

Fiege, C. (2014). *Faire Vergleiche in der Schulleistungsforschung: Methodologische Grundlagen und Anwendung auf Vergleichsarbeiten* (Dissertation). Friedrich-Schiller-Universität Jena. Verfügbar unter http://uri.gbv.de/document/gvk:ppn:776690906

Fiege, C. (2016). Faire Vergleiche bei Vergleichsarbeiten: Möglichkeiten und Grenzen. In B. Groot-Wilken, K. Isaac & J.-P. Schräpler (Hrsg.), *Sozialindices für Schulen: Hintergründe, Methoden und Anwendung* (S. 71–96). Waxmann.

Fiege, C., Reuther, F. & Nachtigall, C. (2011). Faire Vergleiche? Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung, 1*(2), 133–149. https://doi.org/10.1007/s35834-011-0009-x

Fischer, F. T., Schult, J. & Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education, 28*(2), 529–543. https://doi.org/10.1007/s10212-012-0127-4

Groß Ophoff, J., Koch, U. & Hosenfeld, I. (2019). Vergleichsarbeiten in der Grundschule von 2004 bis 2015. In J. Zuber, M. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (S. 205–228). Springer VS. https://doi.org/10.1007/978-3-658-22241-3_9

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Isaac, K. (2008). *Handreichung zum fairen Vergleich bei VERA*. Universität Koblenz-Landau. Verfügbar unter https://dev-schulentwicklung.qua-lis.de/e/upload/images/VERA_Handreichung_fairer_Vergleich_2010.pdf

Isaac, K. & Hosenfeld, I. (2008). Faire Ergebnisrückmeldungen bei Vergleichsarbeiten. In J. Ramseger & M. Wagener (Hrsg.), *Chancenungleichheit in der Grundschule: Ursachen und Wege aus der Krise* (S. 143–146). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91108-3_22

Kemethofer, D. & Wiesner, C. (2019). Verändern Bildungsstandards, Standardüberprüfungen und Ergebnisrückmeldungen die schulische Arbeit? Wahrnehmung, Rezeption und Nutzung aus Perspektive der Schulaufsicht. In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (S. 229–243). Springer VS. https://doi.org/10.1007/978-3-658-22241-3_10

Kempert, S., Edele, A., Rauch, D., Wolf, K.M., Paetsch, J., Darsow, A. et al. (2016). Die Rolle der Sprache für zuwanderungsbezogene Ungleichheiten im Bildungserfolg. In C. Diehl, D. Hunkler & C. Kristen (Hrsg.), *Ethnische Ungleichheiten im Bildungsverlauf* (S. 157–241). Springer VS. https://doi.org/10.1007/978-3-658-04322-3_5

Koch, U. (2011). *Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten? Datenkompetenz von Lehrkräften und die Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten*. Waxmann.

Korngiebel, J. (2014). *Vergleichsarbeiten und ihr Potenzial für die Schul- und Unterrichtsentwicklung: Eine qualitative Untersuchung zur Nutzung der Lernstandserhebungen an hessischen Gymnasien* (Dissertation). Philipps-Universität Marburg. https://doi.org/10.17192/z2014.0221

Kuhl, P., Lenkeit, J., Pant, H.A. & Wendt, W. (2009). Die Kontextuierung von Leistungswerten bei Vergleichs- und Prüfungsarbeiten. Verschiedene Wege, die Zusammensetzung der Schülerschaft in den Rückmeldungen an Schulen und die Schulinspektion zu berücksichtigen. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektion in Deutschland. Eine Zwischenbilanz aus empirischer Sicht* (S. 237–260). Waxmann.

Kultusministerkonferenz. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. KMK.

Leckie, G. & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal, 43*(2), 193–212. https://doi.org/10.1002/berj.3264

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2008). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (Zeitschrift für Erziehungswissenschaft, Sonderheft 8, S. 149–167). https://doi.org/10.1007/978-3-531-90865-6_9

Maier, U., Bohl, T., Kleinknecht, M. & Metz, K. (2011). Einflüsse von Merkmalen des Testsystems und Schulkontextfaktoren auf die Akzeptanz und Rezeption von zentralen Testrückmeldungen durch Lehrkräfte. *Journal for Educational Research Online, 3*(2), 62–93. https://www.waxmann.com/artikelART102690

Nachtigall, C. & Kröhne, U. (2006). Methodische Anforderungen an schulische Leistungsmessung – auf dem Weg zu fairen Vergleichen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen* (S. 59–74). Waxmann.

Pham, G., Freunberger, R., Robitzsch, A., Itzlinger-Bruneforth, U. & Bruneforth, M. (2016). Reliabilität und Stabilität des Index der sozialen Benachteiligung und Kompositionseffekt der Schulen. *Zeitschrift für Bildungsforschung, 6*(3), 345–364. https://doi.org/10.1007/s35834-016-0164-1

Pham, G. & Robitzsch, A. (2014). *„Fairer Vergleich": Technische Dokumentation – BIST-Ü Englisch, 8. Schulstufe, 2013*. BIFIE I Department Bildungsstandards & Internationale Assessments (BISTA). Verfügbar unter https://www.bifie.at/wp-content/uploads/2017/05/TD_Fairer_Vergleich_E8.pdf

Pham, G., Robitzsch, A., George, A.C. & Freundberger, R. (2016). Fairer Vergleich in der Rückmeldung. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit*

R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung (S. 295–332). Facultas.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Schult, J. (2020). *Wie viele Schulen braucht ein fairer Vergleich? Sozialindex-basierte Adjustierungsstrategien im Vergleich*. PsyArXiv. https://doi.org/10.31234/osf.io/9nq2d

Schult, J., Mahler, N., Fauth, B. & Lindner, M.A. (2022). Did students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave. *School Effectiveness and School Improvement, 33*(4), 544–563. https://doi.org/10.1080/09243453.2022.2061014

Schulte, K., Hartig, J. & Pietsch, M. (2016). Berechnung und Weiterentwicklung des Sozialindex für Hamburger Schulen. In B. Groot-Wilken, K. Isaac & J.-P. Schräpler (Hrsg.), *Sozialindices für Schulen: Hintergründe, Methoden und Anwendung* (S. 157–171). Waxmann.

Schwippert, K. (2019). Was wird aus den Büchern? Sozialer Hintergrund von Lernenden und Bildungsungleichheit aus Sicht der international vergleichenden Erziehungswissenschaft. *Journal for Educational Research Online, 11*(1), 92–117. https://www.waxmann.com/artikelART102938

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

Tarkian, J., Maritzen, N., Eckert, M. & Thiel, F. (2019). Vergleichsarbeiten (VERA) – Konzeption und Implementation in den 16 Ländern. In F. Thiel, J. Tarkian, E.-M. Lankes, N. Maritzen, T. Riecke-Baulecke & A. Kroupa (Hrsg.), *Datenbasierte Qualitätssicherung und -entwicklung in Schulen: Eine Bestandsaufnahme in den Ländern der Bundesrepublik Deutschland* (S. 41–103). Springer VS. https://doi.org/10.1007/978-3-658-23240-5_4

Televantou, I., Marsh, H.W., Kyriakides, L., Nagengast, B., Fletcher, J. & Malmberg, L.E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement, 26*(1), 75–101. https://doi.org/10.1080/09243453.2013.871302

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Weishaupt, H. (2016). Sozialindex – Ein Instrument zur Gestaltung fairer Vergleiche: Einführung. In B. Groot-Wilken, K. Isaac & J.-P. Schräpler (Hrsg.), *Sozialindices für Schulen: Hintergründe, Methoden und Anwendung* (S. 13–25). Waxmann.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://doi.org/10.1007/978-3-319-24277-4