LÄNGSSCHNITTANALYSEN IN DER SCHULSPORTFORSCHUNG – EIN METHODENVERGLEICH

von Stefan König & Matthias Lindel

ZUSAMMENFASSUNG | In der Schulsportforschung wurde das Thema Längsschnittanalysen aus methodologischer Sicht bis heute kaum diskutiert. Dies erscheint befremdlich, da der Schulsport mit seinen erzieherischen Zielsetzungen an den Wirkungen konkreter Inszenierungen interessiert sein sollte. Der vorliegende Beitrag verfolgt das Ziel, eine Analyse und Bewertung verschiedener Auswertungstechniken von Längsschnitten vorzunehmen. Zunächst werden Längsschnittstudien bezüglich ihrer erkenntnistheoretischen Grundlagen, ihrer zentralen Merkmale sowie ihrer Stärken und Schwächen mit Blick auf die Sportunterrichtsforschung diskutiert. Dem folgt die vergleichende Darstellung und Bewertung eher traditioneller Ansätze, der Varianzanalyse mit Messwiederholung und der Mehrebenenmodellierung. Anschließend werden Wachstumskurvenmodelle, die in der Sportunterrichtsforschung bisher kaum eine Rolle spielen, als gewinnbringende Alternative vorgestellt. Der Beitrag schließt mit einer Zusammenfassung sowie einigen Handlungsempfehlungen für Längsschnittstudien in der Sportunterrichtsforschung.

Schlüsselwörter: Sportunterrichtsforschung, Längsschnittanalysen, Varianzanalysen, Mehrebenenmodelle, Wachstumskurvenmodelle

LONGITUDINAL ANALYSES IN RESEARCH ON TEACHING IN PHYSICAL EDUCATION–A COMPARISON OF METHODS

ABSTRACT | In Research on Teaching in Physical Eduaction the issue of longitudinal data analysis has rarely been discussed until today. This seems to be strange because Physical Education should be interested in the effects of its specific enactments. This article aims at investigating and discussing various techniques of analysis of longitudinal data. Thus, we first discuss longitudinal studies with reference to their epistemological fundament, their central features, as well as their strengths and weaknesses with a view to RT-PE. Second, we compare and assess two traditional approaches, the analysis of variance with repeated measures and multilevel modelling. Third, we present the concept of latent growth curve modelling, a statistical approach that has been applied rather seldom in RT-PE, as a seminal alternative. The article concludes with a summary and some recommendations for conducting longitudinal data analysis.

Key Words: Research on Teaching in Physical Education, longitudinal data analysis, analysis of variance, multilevel modelling, latent growth curve modelling

LÄNGSSCHNITTANALYSEN IN DER SCHULSPORTFORSCHUNG – EIN METHODENVERGLEICH

1 | EINLEITUNG

Wird die Schulsportforschung aus einem methodologischen Blickwinkel betrachtet, dann sind heute unterschiedlichste Zugänge die Regel (Friedrich, 2000; Hemphill et al., 2012; König, 2002; Novak & Bernstein, 2015; Thiele, 2008), um diejenigen sport- und bewegungsthematischen Handlungs- und Erfahrungszusammenhänge zu untersuchen, die Schüler*innen und Lehrkräfte herstellen, gestalten und entwickeln (Balz et al., 2011). Engt man die Perspektive auf die Sportunterrichtsforschung als ein Kerngebiet der Schulsportforschung ein, haben wir es mit einer Domäne der Lehr-Lern-Forschung zu tun, in der unter anderem die Effekte von sportunterrichtlichen Maßnahmen untersucht werden (Silverman & Skonie, 1997; Töpfer et al., 2020; Wolters, 2011). Eine solche Wirkungsforschung, also die Analyse von Unterrichtsthemen und -programmen sowie deren Effekte auf die Erziehung und Bildung von Schüler*innen, stellt innerhalb der Sportunterrichtsforschung einen relevanten Bereich dar, wobei mit Blick auf die in Deutschland dominante Konzeption des erziehenden Sportunterrichts sich Wirkungsforschung sowohl auf persönlichkeitsbildende als auf qualifikatorische Erziehungsziele bezieht – auch wenn diese in der schulsportlichen Praxis stets miteinander verknüpft sind bzw. sein sollten (Prohl, 2010; 2012).

Obwohl die Schulsport- bzw. die Sportunterrichtsforschung in den letzten 20 Jahren Thema mehrerer spezifischer Publikationen sowie einschlägiger Kongresse war (u. a. Aschebrock & Stibbe, 2018; Balz et al., 2011; Brandl-Bredenbeck & Stefanie, 2009; Friedrich, 2002) und zwischenzeitlich eine respektable Anzahl an spezifischen empirischen Studien vorliegt, wurde das Thema Längsschnittanalysen methodologisch und methodisch selten explizit in diesem wissenschaftlichen Kontext diskutiert. Dennoch lassen sich überblicksartig einige Entwicklungen und Tendenzen festhalten:

(1) Empirische Studien zur Überprüfung von Lern- oder Trainingseffekten werden in der Sportunterrichtsforschung zunehmend häufiger durchgeführt, wobei folgende Beobachtungen auffallen (Bähr et al., 2011; König, 2011; Töpfer et al., 2020):

- Viele Untersuchungen bedienen sich quasi-experimenteller Designs, was den Bedingungen des Settings geschuldet ist, da Effekte spezifischer Unterrichtsprogramme analysiert werden sollen.
- Als Datenerhebungsinstrumente kommen häufig sportmotorische Tests, vermehrt Fragebögen und vereinzelt Experten-Ratings zum Einsatz.
- Was das konkrete Setting betrifft, findet die große Mehrzahl dieser Studien im Regelsportunterricht statt, d. h., es stehen gewöhnlich drei Stunden Sport pro Woche für Interventionen zur Verfügung.
- Die Interventionszeiträume variieren in der Regel zwischen einer und 12 Wochen[1], was unterschiedlichen Tempi der Veränderungsverläufe der Zielvariablen geschuldet ist.

1 Untersuchungen mit längeren Interventionszeiträumen finden sich wenig; aktuelle Ausnahmen sind unter anderem Kalnbach (2019), Lindel (2018) und Wirszing (2015).

(2) Die Auswertung erfolgt in den meisten Studien mittels gepaarter *t*-Tests oder Varianzanalysen mit Messwiederholung; multivariate Verfahren, die in anderen Unterrichtswissenschaften (u. a. Lüdtke et al., 2007; Weißeno, 2019) zu beobachten sind, finden sich in der Schulsportforschung noch eher selten. Dennoch kommen in sportpädagogischen Ansätzen zunehmend komplexere Verfahren zur Anwendung, wie beispielsweise Mehrebenenmodelle und ihre Potenziale zur Erforschung von Unterrichtsprozessen (König, 2019a; Wirszing, 2015), latente Profilanalysen im Rahmen der Vermittlung von Gesundheitskompetenz (Schmid et al., 2020) oder MANOVAs[2] mit Messwiederholung (Kalaja et al., 2012).

Auffallend ist, dass Limitationen und Potenziale verschiedener Auswertungsverfahren für Längsschnittdaten in der sportpädagogischen Literatur bisher kaum systematisch diskutiert wurden. Dies erscheint aber notwendig, da der Schulsport mit seinen erzieherischen Zielsetzungen an den Wirkungen seiner konkreten Inszenierungen interessiert sein sollte (Bräutigam, 2008; Prohl, 2010). Wird dieser Gedanke präzisiert, dann sollte Schulsport- bzw. Sportunterrichtsforschung nicht nur darauf abzielen, Erkenntnisse über seine Akteure, über seine Strukturen und deren Rahmenbedingungen zu generieren, sondern sich auch verstärkt mit spezifischen Wirkungen „über die Zeit" auseinanderzusetzen (Bräutigam, 2008; Wolters, 2011). Insofern ist es angezeigt, den Begriff „Veränderung über die Zeit", die damit verknüpften Forschungsinteressen sowie zentrale Merkmale und angemessene Auswertungstechniken für Längsschnittdaten zu klären.

Vor diesem Hintergrund zielt der vorliegende Beitrag darauf ab, eine vergleichende Analyse und Bewertung verschiedener Auswertungstechniken von Längsschnittdaten vorzunehmen. Hierzu werden in einem ersten Abschnitt Längsschnittstudien bezüglich ihrer erkenntnistheoretischen Grundlagen, ihrer zentralen Merkmale sowie ihrer Stärken und Schwächen mit Blick auf die Schulsport- bzw. die Sportunterrichtsforschung betrachtet und diskutiert. Dem folgt in einem zweiten Schritt die vergleichende Darstellung und Bewertung traditioneller Ansätze, also der Varianzanalyse mit Messwiederholung und der Mehrebenenmodellierung. In Kapitel 4 werden Wachstumskurvenmodelle, die in der Sportunterrichtsforschung bisher kaum eine Rolle spielen, als eine Alternative vorgestellt und beschrieben. In beiden Kapiteln wird bei der Beschreibung und Bewertung der genannten Auswertungsverfahren auf der Basis eines fiktiven Datensatzes erläutert, welche Möglichkeiten, Grenzen und Schwierigkeiten die jeweiligen Ansätze haben. Eine Zusammenfassung sowie Handlungsempfehlungen für Längsschnittstudien in der Sportunterrichts- und Schulsportforschung und deren statistische Auswertung runden den Beitrag ab.

2 | LÄNGSSCHNITTSTUDIEN IN DER SPORTUNTERRICHTFORSCHUNG

Unabhängig vom inhaltlichen Fokus sind viele Untersuchungen in der Sportunterrichtsforschung als Längsschnitte angelegt – das heißt, Daten werden zu mehreren Messzeitpunkten an denselben Teilnehmerinnen und Teilnehmern erhoben (Diekmann, 2011; Fitzmaurice et al. 2011; Singer &

2 MANOVA = *multivariate Varianzanalyse*; hierbei handelt es sich um Varianzanalysen mit mehreren abhängigen Variablen.

Willett, 2003); dieses Vorgehen wird auch als Panelstudie bezeichnet[3]. Generelles Ziel dieser Studien ist, intraindividuelle Veränderungen („within person") und interindividuelle Unterschiede („between person") statistisch zu modellieren, um Heterogenität (Varianz) zwischen Individuen über die Zeit abbilden zu können; dabei spielt es an dieser Stelle (noch) keine Rolle, ob eine rein längsschnittliche Entwicklung oder ein experimenteller Ansatz im Fokus steht. Statistisch betrachtet werden Innersubjektfaktoren, Zwischensubjektfaktoren und deren Interaktionen analysiert, wofür die schließende Statistik verschiedene Verfahren anbietet, die unterschiedliche Voraussetzungen und Gegebenheiten einfordern sowie spezifische Stärken und Schwächen aufweisen (Field, 2014; Geiser, 2010; Hox, 2010).

Im folgenden Abschnitt werden zentrale Merkmale von Längsschnittstudien dargestellt, bevor diese auf die Sportunterrichtsforschung und ihre Besonderheiten und Erkenntnisinteressen übertragen werden.

2.1 | TERMINOLOGISCHE UND ERKENNTNISTHEORETISCHE GRUNDLAGEN

Im Gegensatz zu Querschnittsstudien, bei denen eine oder mehrere empirische Beobachtungen einmalig an den Teilnehmer*innen vorgenommen werden (Johnson & Christensen, 2014, S. 403), bestehen Längsschnittstudien aus empirischen Beobachtungen, die mehrmals hintereinander mit denselben Personen durchgeführt werden, um Veränderungen über die Zeit („change over time") zu modellieren (Conzelmann et al., 2013, S. 321). Für die Sportunterrichtsforschung ist an dieser Stelle zunächst zu klären, auf welche Ausprägung von Zeit sich die Aussage bezieht, da aus einer unterrichtswissenschaftlichen Perspektive Zeithorizonte theoretisch von wenigen Stunden bzw. Tagen bis hin zu Jahren denkbar sind (Keller, 2004; Luke, 2004). Folgende idealtypischen Zeithorizonte sind, auch mit Blick auf den aktuellen Forschungsstand der Sportunterrichtsforschung, relevant:

- *Kurzfristige Szenarien* beziehen sich auf eine oder mehrere Sportstunden (Lutz, 2018) oder auf eine kompakte schulsportliche Maßnahme, wie z. B. eine Wintersportwoche oder Projekttage (Oesterhelt, 2011). Sie zielen darauf ab, Prozesse zu untersuchen, die in sportlichen Kontexten eher schnell ablaufen können, also motorische Prozesse des Neulernens, kognitive Lern- oder Gruppenbildungsprozesse.
- Ein *mittelfristiges Szenario* bezieht sich auf Veränderungen im Laufe von mehreren Wochen und stellt das am häufigsten verwendete Zeitmodell in der Sportunterrichtsforschung dar; es findet insbesondere Anwendung, wenn es um motorische Lern- oder Trainingsprozesse geht (König, 2016; Schiemann & Pargätzi, 2016; Thienes, 2016). Allein aufgrund physiologischer Gegebenheiten sind gewisse Mindestanforderungen an die Zeit notwendig, welche für den Sportunterricht mit einem Trigger-Wert von sechs Wochen abgebildet werden konnten (König, 2011).
- *Langfristige Szenarien* umfassen einen Zeithorizont von Monaten bzw. einem oder mehreren Schuljahren und sind in der Sportunterrichtsforschung eher selten. Da Lern- und Trainings-

3 Im Gegensatz zu *Panelstudien* bilden *Trendstudien* auch Längsschnitte ab, allerdings werden Daten zu mehreren Messzeitpunkten an unterschiedlichen Stichproben erhoben.

prozesse im Sport längerfristig anzulegen sind, besonders dann, wenn stabile Effekte gewünscht werden, sind solche Zeithorizonte für die Schulsportforschung notwendig und interessant, was etwa Wartenberg et al. (2014) für die schulische Entwicklung von Schülerathlet*innen, Wirszing (2015) für die motorische Entwicklung von Grundschulkindern oder Kalnbach (2019) für die Entwicklung der motorischen Leistungsfähigkeit von Sekundarstufenschüler*innen zeigen konnten.

Neben den beschriebenen Zeithorizonten sind es aber auch die erkenntnistheoretischen Interessen, die eine weitere Differenzierung der eingangs genannten Zielsetzung erforderlich machen. Bezüglich der von uns gewählten Thematik des Längsschnitts sind dies die folgenden Forschungsinteressen, die zunächst in Reinform gegenübergestellt werden:

- *Interventionsstudien* zeichnen sich dadurch aus, dass sie die Wirkung eines spezifischen Programms prüfen und folglich in der Regel mit experimentellen oder quasi-experimentellen Designs arbeiten (Twisk, 2010). Sie sind primär an Veränderungen von Messzeitpunkt 1 nach 2 bzw. am Output von Unterrichtsprogrammen interessiert, was sich in der Sportunterrichtsforschung typischerweise bei Trainingsexperimenten, Studien zur Förderung des Selbstkonzepts oder Wirksamkeitsprüfungen von Lehrmethoden anbietet; hier gilt das 2 x 2-Design als Klassiker, welches je nach Forschungsinteresse verändert oder erweitert werden kann (Lindel, 2018; Maxwell & Delaney, 2004; Twisk, 2010).
- *Entwicklungsstudien* hingegen beobachten, beschreiben und erklären Veränderungen ohne spezifische Interventionen über längere Zeiträume (Meinel & Schnabel, 2007; Wartenberg et al., 2014) und gelten als klassische Längsschnittstudien. Allerdings geht es auch ihnen darum, Variablen in Abhängigkeit von Prädiktoren zu untersuchen, um deren Einflüsse auf die Entwicklung von Individuen abbilden zu können.

Für die in diesem Beitrag diskutierte Thematik ist entscheidend, dass wir uns von einem reinen „Entweder-oder-Denken" verabschieden, da beide Ansätze miteinander kombiniert werden können bzw. gerade in der Unterrichtsforschung bereits kombiniert wurden. Beispiele hierfür sind Studien, die die Wirkung einer Intervention sowie deren Stabilität prüfen oder nach einer Phase ohne Treatment eine Intervention in einen Längsschnitt integrieren. Unabhängig von der konkreten Struktur einer Studie müssen alle Untersuchungen, die eine Veränderung über die Zeit analysieren, drei Strukturmerkmale aufweisen, welche im folgenden Abschnitt diskutiert werden.

2.2 | MERKMALE VON LÄNGSSCHNITTSTUDIEN

Längsschnittstudien erheben unabhängig vom zeitlichen Horizont und dem jeweiligen Forschungsinteresse mehrfach Daten an einer Kohorte (Bortz & Döring, 2006; Schnell et al., 2011). Ziel von Längsschnittstudien ist es daher, herauszuarbeiten, wie sich Individuen einer Stichprobe über die Zeit verändern, wobei unabhängige Variablen diese Veränderung erklären sollen. Für die Planung solcher Forschungsansätze werden in der einschlägigen Literatur drei obligatorische Strukturmerkmale genannt (Fitzmaurice et al., 2011; Singer & Willett, 2003):

(1) *Eine angemessene Anzahl an Messzeitpunkten*. Während für rein experimentelle Designs zwei Messzeitpunkte genügen, benötigen Längsschnittstudien drei oder mehr Wellen an Datener-

hebungen, da ansonsten individuelle Veränderungen nicht robust abgebildet werden, d. h. verzerrt sein können. Allerdings kann durchaus auch für experimentelle Designs ein dritter Messzeitpunkt von Interesse sein, um entweder die Form der Veränderung oder aber ihre Stabilität zu prüfen. Der häufig verbreitete Ansatz, mit zwei Messungen zu arbeiten, ist – von reinen Experimenten abgesehen – aus verschiedenen Gründen fragwürdig:

- Zwei Messzeitpunkte können eine individuelle Entwicklungslinie nicht gut abbilden, da Veränderung als „… the simple difference between scores assessed on two measurement occasions" betrachtet wird, d. h. „change is regarded as the acquisition or loss of the focal increment" (Singer & Willett, 2003, S. 10), e.g. an achievement or an attitude. Damit sind Aussagen über die Form der Veränderung oder über kritische Zeitpunkte eventueller Veränderungen nicht möglich.
- Zwei Messzeitpunkte können keine Messfehler berücksichtigen, eine Unterscheidung von wahrer Entwicklung und Messfehler ist nicht möglich. Dies ist besonders dann kritisch, wenn beispielsweise Ergebnisse im Prätest zu niedrig und im Posttest zu hoch sind und es damit zu teilweise extremen Verzerrungen statistischer Schätzungen kommt.

Insofern ist bei Längsschnittstudien – trotz eines gerade auch für den Sportunterricht größeren Aufwandes – von einem Design mit zwei Messzeitpunkten abzuraten, außer die Forschungsfrage verlangt die Umsetzung eines experimentellen bzw. quasi-experimentellen Designs unter möglichst standardisierten Bedingungen.

(2) *Ein Messergebnis (AV), das sich systematisch über die Zeit verändert bzw. zumindest theoretisch verändern kann*. Zieht man bezüglich dieses Axioms das Angebots-Nutzungs-Modell von Helmke (2010) heran, dann wird deutlich, dass Lernaktivitäten, die durch Unterrichtsangebote von Lehrenden initiiert werden, Wirkungen erwarten lassen. Dies konnte auch für den Sportunterricht und für außerunterrichtliche Schulsportangebote in vielen Lernbereichen gezeigt werden. Exemplarisch soll diese Aussage wie folgt belegt werden:

- Betrachtet man den Bereich der Studien zu konditionellen und koordinativen Trainingsprozessen, so wurde in einer Vielzahl von Studien deutlich, dass im Schulsport eine abhängige Variable, in diesem Fall eine motorische Fähigkeit, verbessert werden kann.
- Etwas bescheidener ist die Studienlage im Bereich des motorischen Lernens (Töpfer et al., 2020). Dennoch kann das eingangs formulierte Axiom auch für diesen Bereich bestätigt werden, Gleiches gilt für taktische Lernprozesse (Memmert & König, 2007).
- Veränderungen über die Zeit konnten auch für personale und soziale Aspekte empirisch belegt werden, so etwa für das soziale Selbstkonzept (Magnaguagno et al., 2016) oder für die Gruppenkohäsion (Oesterhelt, 2011).

Für die Sportunterrichtsforschung kann somit festgehalten werden, dass Längsschnittanalysen die Veränderung von Variablen erklären, die aus den Zielsetzungen eines erziehenden Sportunterrichts abgeleitet werden und (zumindest theoretisch) veränderbar sein müssen.

(3) *Einen vernünftigen Messzeittakt, der eine inhaltliche Begründung erfordert.* Werden Längsschnittstudien im Überblick betrachtet, so fällt auf, dass unterschiedliche Messzeittakte vorliegen.

Sie reichen von mehreren Messungen am Tag (z. B. Wirkungsstudien mit Medikamenten) bis hin zu mehreren Jahren (z. B. Entwicklung von Aktivitäten des täglichen Lebens). Ein Messzeittakt in der Sportunterrichtsforschung muss sich am jeweiligen Forschungsinteresse orientieren, da abhängige Variablen im Sport sich nach unterschiedlichen Zeitverläufen verändern. Dies soll anhand folgender Überlegungen erläutert werden:

- Grundsätzlich ist davon auszugehen, dass im Schulsport eher selten kurze Messzeittakte angemessen sind, da Lernprozesse in der Regel eine gewisse Zeit beanspruchen. Dennoch können sich koordinative und kognitive Veränderungen kurzfristig einstellen (Lutz, 2018), was theoretisch eine Messung im Wochenrhythmus erfordert. Kurze Messzeittakte stellen allerdings ein Problem in der Realisierung vor Ort dar.
- Was eine Verbesserung von motorischen Fähigkeiten angeht, so sind Messzeit-Takte im Mehrwochen-Rhythmus angemessen (König, 2011; Lindel, 2018), da sich in der Regel vor einem Zeitraum von 5 bis 8 Wochen keine Veränderungen nachweisen lassen; Gleiches gilt für Einstellungen und Haltungen (Magnaguagno et al., 2016).
- Schließlich kann auch ein langfristiger Messzeittakt notwendig sein, wenn motorische Entwicklungsparameter, bspw. das Gleichgewichtsvermögen (Hirtz & Forschungszirkel „N. A. Bernstein“, 2007), gemessen werden; in solchen Fällen sind Messungen alle paar Monate angezeigt, vor dem Hintergrund der relevanten Prozesse ist dies aber gerechtfertigt.

Was das dritte Axiom betrifft, kann festgestellt werden, dass Aktivitäten im Sportunterricht in unterschiedlichen Zeittakten ablaufen. Längsschnittstudien müssen dies berücksichtigen und folglich ihren Takt im Einzelfall jeweils inhaltlich begründen.

3 | TRADITIONELLE MODELLIERUNGSANSÄTZE FÜR LÄNGSSCHNITTE

Werden Standardwerke zur empirisch-quantitativen Erforschung von Veränderungen über die Zeit (Fitzmaurice et al., 2011; Hox, 2010; Maxwell & Delaney, 2004; Rasch et al., 2014; Singer & Willett, 2003) konsultiert, dann findet man häufig zwei Modellierungsansätze, die zur Analyse von Längsschnittdaten eingesetzt werden: Es sind dies die Varianzanalyse mit Messwiederholung, die als Klassiker für diese Modellierung gilt, sowie die Mehrebenenmodellierung für Längsschnittdaten, die den für die Unterrichtsforschung typischen, hierarchisch strukturierten Datensätzen[4] gerecht wird. Sie werden in den beiden folgenden Abschnitten beschrieben, diskutiert und bewertet; ebenfalls wird auf jeweils spezifische Softwarepakete eingegangen, ohne einen Anspruch auf Vollständigkeit zu haben. Zum Zwecke der Anschaulichkeit erfolgt dies mittels eines fiktiven Datensatzes, der im Folgenden vorgestellt wird:

- Der Datensatz besteht aus n = 800 Schüler*innen, die im Schnitt 14 Jahre alt sind (SD: 2,63); die Gesamtstichprobe setzt sich aus n = 380 Mädchen und n = 420 Jungen zusammen. 411 der Proband*innen sind Mitglied in einem Sportverein, 389 geben an, dies nicht zu sein.

4 Von hierarchisch strukturierten Datensätzen wird dann gesprochen, wenn sich Beobachtungen eines Datensatzes hierarchisch übergeordneten Einheiten zuordnen lassen, z. B. wenn Messungen an Schüler*innen (Level 1) vorgenommen werden, die in Klassen (Level 2) gruppiert sind; die Klassen ihrerseits sind wiederum in Schulen (Level 3) gruppiert.

- Die Schüler*innen kommen aus insgesamt 31 Klassen verschiedener Schulen, was in diesem Gedankenexperiment den Kontexteffekt darstellt. An dieser Stelle wird bereits die hierarchische Schachtelung der Daten erkennbar. Die vier Messzeitpunkte (Level 1) sind in den jeweiligen Schülerinnen und Schülern gruppiert (Level 2), die wiederum in ihren jeweiligen Klassen (Level 3) geschachtelt sind. Zu erwähnen ist, dass der zuletzt genannte Wert der Untergrenze für die benötigte Anzahl an Kontexteinheiten auf Level 2 bzw. Level 3 entspricht, da in der Regel mindestens 30, eher 50 übergeordnete Einheiten gefordert werden (Hox, 2010).
- Ziel der hypothetischen Untersuchung ist, exemplarisch die Wirkungen von Krafttraining über vier Messzeitpunkte zu analysieren. Das Beispiel sieht weiterhin vor, dass zu jedem Messzeitpunkt vier unterschiedliche spezifische sportmotorische Tests eingesetzt werden.
- Ein weiteres Merkmal des fiktiven Datensatzes ist, dass n = 400 Schüler*innen der Experimentalgruppe und die gleiche Anzahl der Kontrollgruppe angehören. Insofern entspricht der Ansatz dem in der Sportunterrichtsforschung häufig verwendeten quasi-experimentellen Design mit Cluster-Randomisierung (Johnson & Christensen, 2014, S. 357f.).
- Weiterhin ist anzumerken, dass die vier Messzeitpunkte nicht äquidistant[5] sind, was in vielen Situationen des Schulsports wegen Stundenplanung, Ferieneinschnitten oder anderer schulischer Aktivitäten zutrifft.
- Schließlich weist der Datensatz im Schnitt etwa 10 % fehlende Werte pro abhängiger Variable auf, für die allerdings das Merkmal „missing completely at random“[6] gilt. Dies ist für die Sportunterrichtsforschung ein normaler Wert, führt aber dazu, dass im schlechtesten Fall und aufgrund der Prozedur „listwise deletion“ (alle Datensätze mit einem oder mehreren fehlenden Werten werden von der statistischen Berechnung ausgeschlossen) in unserem Fall mit nur 256 vollständigen Datensätzen gearbeitet werden kann, was die Testpower gegebenenfalls verringert und damit die Gefahr erhöht, die Nullhypothese abzulehnen (Johnson & Christensen, 2014, S. 571).

Auf weitere deskriptive Koeffizienten wird an dieser Stelle verzichtet, da sie für die folgenden methodologischen Überlegungen nicht relevant sind. Abschließend möchten wir nochmals darauf hinweisen, dass die abhängigen Variablen Kraft 1 bis 4 nur Platzhalter für alle möglichen Zielsetzungen des Sportunterrichts darstellen.

3.1 | VARIANZANALYSE MIT MESSWIEDERHOLUNG (rANOVA)

Zu den bekanntesten statistischen Analyseverfahren für die Auswertung von Längsschnittstudien (mit parametrischem Datenniveau) gehört neben dem *t*-Test die Varianzanalyse, die als Erweiterung des *t*-Tests für mehr als zwei Messzeitpunkte und/oder mehr als zwei Gruppen betrachtet werden kann (Bortz & Schuster, 2010). Da die Varianzanalyse in den meisten Lehrbüchern

5 Nicht äquidistante Messzeitpunkte bedeuten, dass der zeitliche Abstand zwischen den Messzeitpunkten unterschiedlich groß ist.

6 „Missing completely at random“ bedeutet, dass fehlende Werte keinerlei Systematik im Verhältnis zum Gesamtdatensatz aufweisen und zudem kein Zusammenhang zwischen ihnen und den Werten der jeweiligen Variable besteht (Allison, 2002, S. 3f.).

zur Einführung in die Statistik behandelt wird (Raithel, 2008; Rasch et al., 2014; Sedlmeier & Renkewitz, 2013), ist es nicht verwunderlich, dass sie auch in der Sportunterrichtsforschung regelmäßig zur Analyse von Längsschnittdaten eingesetzt wird. Eine rANOVA ermöglicht es, alle drei interessierenden und im Folgenden beschriebenen Effekte bei der Analyse von Längsschnittdaten auf statistische Signifikanz und in Bezug auf die jeweilige Effektgröße zu prüfen:

Zunächst kann mithilfe der (einfaktoriellen) Varianzanalyse untersucht werden, ob sich das T_0-Niveau der Experimentalgruppe von dem der Kontrollgruppe signifikant unterscheidet[7], indem die zu überprüfende Variable als abhängige Variable und die Gruppierungsvariable als Faktor in der jeweils eingesetzten Statistiksoftware definiert wird. Der Innersubjekteffekt der Experimentalgruppe kann im Anschluss analysiert werden, indem die rANOVA nur auf den Teildatensatz der Experimentalgruppe angewendet wird. Als Innersubjektfaktor, der aufgrund der Anzahl der Messzeitpunkte über die entsprechend korrespondierende Anzahl an Stufen verfügt – im vorliegende Fall folglich vier Stufen – wird die zu T_0 bis T_3 erhobene Variable definiert. Als Zwischensubjektfaktor, der aufgrund der zuvor vorgenommen Fallauswahl lediglich die Daten der Experimentalgruppe enthält, wird die Gruppierungsvariable definiert. Mithilfe des Haupteffekts des messwiederholten Innersubjektfaktors, der bei diesem Vorgehen den Haupteffekt der Zeit darstellt, kann geschätzt werden, ob sich die Experimentalgruppe über die Zeit von T_0 zu T_3 in Bezug auf die erhobene Variable signifikant verändert hat und wie groß dieser Effekt ist. Die Prüfung des Innersubjekteffekts der Experimentalgruppe sollte als zweiter Analyseschritt nach dem Vergleich des Eingangsniveaus zwischen der Experimental- und der Kontrollgruppe grundsätzlich immer durchgeführt werden, da sonst eine eindeutige Bewertung der Intervention nicht unter allen Umständen möglich ist.

Ein einfaches Beispiel verdeutlicht dies: Verbessert sich die Experimentalgruppe von T_0 zu T_3 lediglich deskriptiv, jedoch nicht signifikant, und die Kontrollgruppe verschlechtert sich von T_0 zu T_3 signifikant, dann zeigt sich der in Abbildung 1 dargestellte Entwicklungsverlauf der beiden Gruppen. Fällt der Interaktionseffekt Zeit*Gruppe signifikant aus, kann der Eindruck und daraus die falsche Interpretation entstehen, dass das zu untersuchende Interventionsprogramm erfolgreich ist, da die Experimentalgruppe eine Steigerung ihrer Leistung erzielt hat, die sich zudem aufgrund des getesteten Interaktionseffekts Zeit*Gruppe signifikant von der Entwicklung der Kontrollgruppe unterscheidet. Dabei wird jedoch übersehen, dass sich die Experimentalgruppe über die Zeit hinweg gar nicht signifikant verändert hat; die Aussage, das Interventionsprogramm sei im Sinne einer Leistungssteigerung[8] erfolgreich, ist damit hinfällig.

7 Angemerkt sei an dieser Stelle, dass dieses Vorgehen, auch wenn es in der wissenschaftlichen Praxis oft so durchgeführt wird, aufgrund von forschungstheoretischen und -praktischen Gründen problematisch ist und besser auf Verfahren der Äquivalenztestung beim Nachweis eines vergleichbaren T_0-Niveaus zurückgegriffen werden sollte (Lindel, 2018; Walker & Nowacki, 2011).

8 An dieser Stelle sei angemerkt, dass natürlich auch Interventionsprogramme möglich sind, deren Ziel nicht die Steigerung, sondern die Erhaltung der Leistung über die Zeit ist. Da der Schulsport aber in der Regel eine Leistungssteigerung beabsichtigt, wird das Vorgehen bei der Auswertung solcher Studien hier nicht thematisiert.

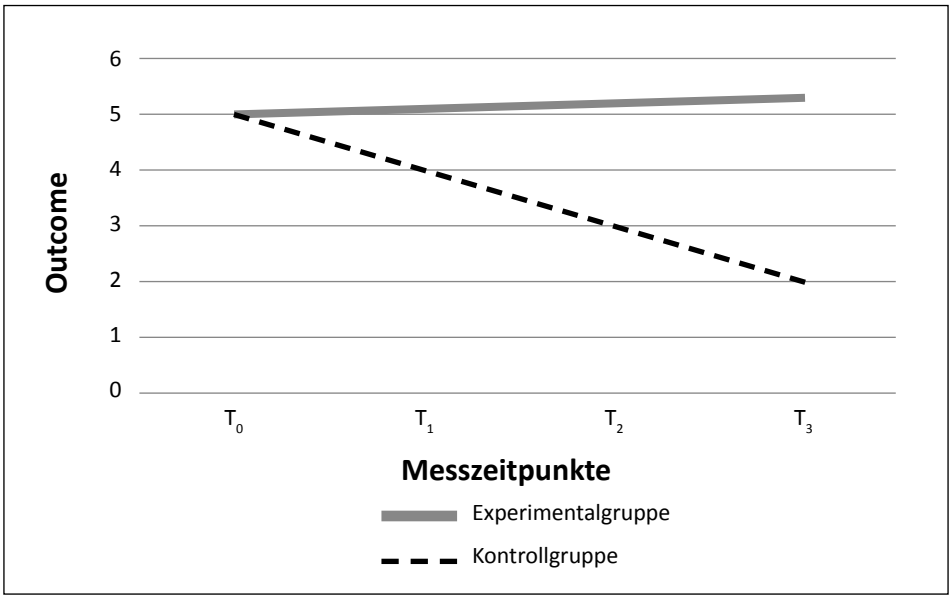

Abb. 1: Mögliche Fehlinterpretation bei Nichtdurchführung der Innersubjekttestung der Experimentalgruppe

Wird für die Analyse des Interaktionseffekts Zeit*Gruppe die im zweiten Schritt vorgenommene Fallauswahl wieder aufgehoben, sodass nun sowohl der Datensatz der Experimental- als auch der Kontrollgruppe in die rANOVA miteinfließen, kann anhand dieses Interaktionseffekts überprüft werden, ob sich die Experimentalgruppe über die Zeit signifikant anders als die Kontrollgruppe entwickelt und wie groß dieser Effekt ist.

Da die rANOVA in den gängigen Statistiksoftwares wie *IBM SPSS*, *Stata* oder *StatSoft STATISTICA* implementiert ist, kann sie in der Regel auch von Neulingen relativ problemlos im Bereich der statistischen Analyse eingesetzt werden. Mit der Varianzanalyse (mit Messwiederholung) können alle drei interessierenden Effekte (Vergleich des T_0-Niveaus, Haupteffekt für die Experimentalgruppe sowie der Interaktionseffekt Zeit*Gruppe) hinsichtlich Signifikanz und Effektgröße berechnet werden. Darüber hinaus ist es bei Anwendung der rANOVA möglich, das bisher vorgestellte Analysemodell bezüglich des möglichen Einflusses weiterer Variablen, z. B. des Geschlechts, auf die Entwicklung der Leistung über die Zeit zu erweitern beziehungsweise deren Effekte zu analysieren und zu kontrollieren. Ebenfalls bietet sich bei der rANOVA die Möglichkeit, Kovariate in das Modell miteinzubeziehen, um mittels dieser Kontrollvariablen mögliche Einflüsse auf das Untersuchungsergebnis zu prüfen (Bortz & Schuster, 2010).

Jedoch weist die Varianzanalyse (mit Messwiederholung) für die Sportunterrichtsforschung drei Nachteile auf, die im Folgenden erläutert werden:

- Die Varianzanalyse setzt bei der Analyse des messwiederholten Faktors äquidistante Abstände zwischen den Messzeitpunkten voraus. Diese gleichen Abstände zwischen den einzelnen Messzeitpunkten können im Rahmen von Sportunterrichtsforschung oftmals aus untersuchungspraktischen Gründen, beispielsweise wenn aufgrund von Ferieneinschnitten die Zeit zwischen dem ersten und dem zweiten Messzeitpunkt fünf Wochen und zwischen dem zweiten und dritten hingegen sieben Wochen beträgt, nicht realisiert werden. Dadurch kommt es zu einer Verletzung der Annahmen der rANOVA.
- Ein zweiter Nachteil der rANOVA ist, dass sie über alle Messzeitpunkte hinweg vollständige Datensätze für die Analyse benötigt. Fehlt gerade bei Untersuchungsdesigns mit mehr als zwei Messzeitpunkten für einen Untersuchungsteilnehmer auch nur ein einziger Wert für einen Messzeitpunkt, wird dieser Untersuchungsteilnehmer komplett aus der statistischen Analyse ausgeschlossen, was dazu führt, dass „missing data lead to more data being deleted" (Field, 2014, S. 818). Infolgedessen kann es durch den Verlust an statistischer Power zu einem erhöhten Betafehlerrisiko kommen (Bortz & Schuster, 2010).
- Ebenfalls ein Nachteil ist, dass beim Einsatz der Varianzanalyse (mit Messwiederholung) im schulischen Kontext diese nicht in der Lage ist, stochastisch abhängige Stichproben, d. h., sich gegenseitig beeinflussenden Messungen[9], korrekt zu schätzen. Das kann zu einer entsprechenden Verzerrung bei der Schätzung der *p*-Werte und weiterer Koeffizienten sowie zu einem ökologischen Fehlschluss und, damit verbunden, zu falschen Interpretationen bei der Beurteilung der jeweiligen Intervention bzw. der zeitlichen Entwicklung führen (Lindel, 2018, S. 83–92).
- Schließlich ist noch herauszuheben, dass eine rANOVA zwischen inter- und intraindividuellen Unterschieden nicht differenzieren kann, da individuelle Werte zu einem Gesamtmittelwert zusammengefasst werden. Gerade eine solche Unterscheidung könnte aber für Untersuchungen im Sportunterricht relevant sein, nämlich dergestalt, dass nach einheitlichen oder unterschiedlichen Verläufen über die Zeit gesucht wird, wenn sich beispielsweise ein Interventionsprogramm auf die teilnehmenden Schüler*innen unterschiedlich auswirkt: Verbessert sich etwa ein Teil der Schüler*innen in ihren Leistungen und ein anderer verschlechtert sich, werden durch die Aggregation der Einzelwerte zu einem Gruppenwert diese unterschiedlichen Entwicklungen unter Umständen übersehen. Für solche Analysen wird die Modellierung individueller Wachstumskurven notwendig (Eid et al., 2013; Schendera, 2008).

Fazit: Unter Bezugnahme auf den eingangs vorgestellten fiktiven Datensatz ermöglicht eine rANOVA die Analyse bzw. Beantwortung folgender Aspekte und Fragen:

(1) Vergleich des Eingangsniveaus der Experimental- und der Kontrollgruppe: Gibt es Unterschiede zwischen den beiden Gruppen in Bezug auf die Kraftleistung in den vier eingesetzten Tests zum ersten Messzeitpunkt und, wenn ja, wie groß sind diese Unterschiede?

9 Im vorliegenden Fall kann die Gruppierung der Schüler*innen in Klassen dafür sorgen, dass die Messwerte der einzelnen Schüler*innen einer Klasse aufgrund von Kontexteffekten untereinander ähnlicher sind als im Vergleich zu den Messwerten von Schüler*innen anderer Klassen; dieser Tatbestand wird als „stochastisch abhängig" bezeichnet. Die rANOVA setzt jedoch stochastisch unabhängige Stichproben voraus.

(2) Analyse der zeitlichen Entwicklung der Experimental- sowie der Kontrollgruppe über alle vier Messzeitpunkte hinweg – oder vereinfacht ausgedrückt: Hat sich die Kraftleistung der Experimental- und die der Kontrollgruppe zwischen den vier Messzeitpunkten signifikant verändert und wie groß ist gegebenenfalls die Veränderung der Kraftleistung?

(3) Analyse des Interaktionseffekt Zeit*Gruppe: Entwickelt sich die Experimentalgruppe über die vier Messezeitpunkte hinweg anders als die Kontrollgruppe in dem Sinne, dass die Verbesserung der Experimentalgruppe überzufällig größer ist als die Verbesserung der Kontrollgruppe?

(4) Problematisch ist, dass die rANOVA nicht in der Lage ist, mit zeitlich nicht äquidistanten Messzeitpunkten umzugehen, wobei dies in der praktischen Umsetzung von Schulsportforschung in der Regel nur sehr schwer zu realisieren ist. Zudem führen Missings zu einem Messzeitpunkt zum Ausschluss aller weiteren erhobenen Daten der jeweiligen Person, was entweder den Einsatz von Imputationsverfahren erforderlich macht oder zu einer reduzierten statistischen Power führen kann.

Zusammenfassend kann an dieser Stelle gesagt werden, dass die rANOVA alle grundlegenden statistischen Analysen bei der Auswertung einer Längsschnittstudie durchführen und somit die Frage nach der Wirksamkeit einer Intervention im Rahmen von Schulsportforschung beantworten kann.

3.2 | MEHREBENENANALYSEN (MLM)

In der bildungswissenschaftlichen Forschung und folglich auch im Rahmen von Sportunterrichtsforschung liegt in der Regel eine hierarchische Datenstruktur vor, die auch als hierarchische Schachtelung, „nested data“ oder Clusterung von Daten bezeichnet wird (Snijders & Bosker, 2012). Diese Schachtelung entsteht dadurch, dass einzelne Schüler*innen (Level 1) jeweils in Klassen (Level 2) und diese wiederum in den jeweiligen Schulen (Level 3) organisiert sind (Field, 2014; Goldstein, 1987). Auch ein typisches Längsschnittdesign im Rahmen von Sportunterrichtsforschung kann als hierarchische Schachtelung von Daten angesehen werden: Die einzelnen Messzeitpunkte (Level 1) sind in den jeweiligen Schüler*innen (Level 2) geclustert, diese wiederum in ihren Klassen (Level 3). (König, 2019b; Onwuegbuzie & Hitchcock, 2015; Raudenbush & Bryk, 2002). Abbildung 2 stellt diesen Ansatz schematisch dar.

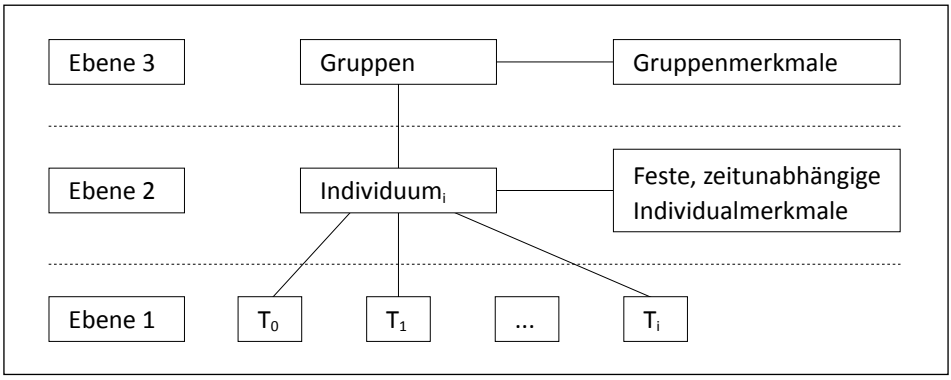

Abb. 2: Mehrebenenstruktur in der Unterrichtsforschung (Field, 2014, S. 817; König, 2019b, S. 111)

Zunächst ist zu klären, ob in dem hier untersuchten Datensatz eine gegebenenfalls vorhandene hierarchische Datenstruktur Auswirkungen auf eine mögliche stochastische Abhängigkeit der Stichprobenelemente hat. Dies erfolgt etwa durch Berechnung des ICC[10] (Hu & Bentler, 1999). Wenn diese Frage nach der entsprechenden Analyse der Daten positiv beantwortet werden muss, kann eine stochastische Unabhängigkeit durch MLM modelliert und adäquat bei der Berechnung der Koeffizienten sowie der zugehörigen *p*-Werte berücksichtigt werden.

Da die rANOVA bei der Berechnung der Koeffizienten sowie der dazugehörigen *p*-Werte davon ausgeht, dass keine Clusterung der Daten vorliegt (Voraussetzung der stochastischen Unabhängigkeit der Stichprobe), werden diese bei Verwendung der rANOVA auf geschachtelte Daten falsch geschätzt, was zu teilweise massiven Fehlinterpretationen führen kann:

„The majority of studies of educational effects – whether classroom experiments, or evaluations of programs, or surveys – have collected and analyzed data in ways that conceal more than they reveal. The established methods [gemeint sind hier in erster Linie die rANOVA bzw. Messverfahren, die auf dem Allgemeinen Linearen Modell basieren, a. d. A.] have generated false conclusion in many studies" (Cronbach et al., 1976, S. 8).

Dass eine solche geschachtelte Datenstruktur im Rahmen von Sportunterrichtsforschung eher die Regel als die Ausnahme ist, zeigt sich beispielsweise in der Untersuchung zu den Wirkungen von DKV-Sound-Karate im Schulsport (Lindel, 2018): Bei acht von neun untersuchten Testitems konnte u. a. mithilfe des ICC die Auswirkung der Schachtelung auf die Datenstruktur nachgewiesen werden, und es zeigten sich unterschiedliche Ergebnisse beim Vergleich des Eingangsniveaus, der Testung des Innersubjekteffekts sowie bei der Testung des Interaktionseffekts. In allen Fällen, bei denen es zu einer Abweichung der Ergebnisse zwischen der rANOVA und der Mehrebenenanalyse (MLM) kam, führte die Anwendung der rANOVA zu statistisch signifikanten Ergebnissen, während die MLM-Berechnung diese aufgrund der korrekten Spezifikation der hierarchischen Datenstruktur verneinte. Darüber hinaus ist MLM flexibler beim Umgang mit unterschiedlichen Abständen zwischen den Messzeitpunkten sowie bei fehlenden Datensätzen:

„Of the advantages of the hierarchical linear model approach to repeated measures, one deserves to be mentioned here. It is the flexibility to deal with unbalanced data structures, for example, repeated measures data with fixed measurement occasions where the data for some (or all) individuals are incomplete or longitudinal data where some or even all individuals are measured at different sets of time points" (Snijders & Bosker, 2012, S. 247).

Beim Einsatz von MLM in Längsschnittstudien werden für alle Ebenen einzelne Regressionsgleichungen aufgestellt, wobei folgende Überlegungen relevant sind:

(1) Die Gleichung für Ebene 1 entspricht der Regressionsgeraden für den Individualwert; sie besteht folglich aus einem Ausgangswert (intercept β_0) und einer Steigung (slope β_1) und entspricht einer linearen Regression.

10 Der Intraclass Correlation Coefficient gibt an, wie viel Einfluss übergeordnete Kontexteinheiten auf die Varianz der abhängigen Variablen haben, und kann daher zur Entscheidungsfindung hinsichtlich der Notwendigkeit einer Mehrebenenanalyse herangezogen werden (Eid et al., 2013).

(2) Die Gleichungen auf den höheren Ebenen entsprechen jeweils den Regressionsgeraden für die Berechnung von intercept (β_0) und slope (β_1) in Abhängigkeit von den jeweiligen Ausprägungen der Variablen der höheren Ebenen.

(3) Abschließend werden alle Einzelgleichungen zu einer Gesamtgleichung zusammengefügt („mixed effect model").[11]

Grundsätzlich können mit dem Verfahren der Mehrebenenanalyse alle interessierenden Analyseebenen (auf Level 1 die Entwicklung über die Zeit, auf Level 2 der Einfluss von Personenmerkmalen und auf Level 3 die Wirkung des Interventionsprogramms oder der Klassenkontext) untersucht und die jeweils relevanten Effekte geschätzt werden.

Die Vorteile von MLM liegen neben der Vermeidung eines ökologischen Fehlschlusses in der Berücksichtigung von hierarchisch strukturierten Daten und damit einer unverzerrten Schätzung von Koeffizienten sowie deren zugehöriger *p*-Werte (Luke, 2004). Darüber hinaus sind Mehrebenenmodelle in der Lage, flexibel hinsichtlich des zeitlichen Abstands der Messzeitpunkte zu sein; nicht äquidistante Messungen können entsprechend innerhalb des Modells auf Level 1 berücksichtigt und somit korrekt modelliert werden. Ebenfalls verbleiben Personen (Level 2) mit einzelnen fehlenden Messungen auf Level 1 in der Untersuchung mit mehr als zwei Messzeitpunkten in der Analyse, da MLM auf der Basis eines (linearen) Regressionsmodells arbeitet. Während fehlende Werte auf Level 1 kein größeres Problem darstellen, sind Missings auf Level 2 oder Level 3 auch bei Verwendung von MLM problematisch, da diese nicht durch das Regressionsmodell geschätzt werden können.

Nachteilig beim Einsatz von MLM ist, dass eine statistische Modellierung deutlich komplexer ist als der Einsatz einer rANOVA, was erweiterte Statistikkenntnisse und eine vertiefte Auseinandersetzung mit den jeweiligen Statistikprogrammen (*HLM, SPSS, STATA, R* …) erfordert. Für die Berechnung von Effektgrößen bei der Verwendung von MLM-Modellen gibt es unterschiedliche Vorgehensweisen, bisher fehlt aber eine standardisierte bzw. durchgängig akzeptierte Berechnungsmethode (Garson, 2013). Auch sind die statistischen Voraussetzungen für MLM-Modelle umfangreicher als bei der Varianzanalyse (mit Messwiederholung), da beispielsweise die Voraussetzungen für Normalverteilung und die Varianz der Residuen nicht mehr nur auf einer, sondern auf mehreren Ebenen untersucht werden müssen (Hox, 2010; Snijders & Bosker, 2012). Zusätzlich steigt der organisatorische Aufwand bei Mehrebenenstudien erheblich an, da ein deutlich größerer Stichprobenumfang benötigt wird (Hox, 2010; Rost, 2013): So werden zwischen 30 und 50 Einheiten auf der obersten Ebene erwartet, was im Falle von Längsschnitten in der Sportunterrichtsforschung 30 bis 50 Klassen entspricht. Im Gegensatz dazu müssen auf der darunterliegenden Ebene – die im beschriebenen Fall Schülerinnen und Schülern entspricht – nicht alle Fälle untersucht werden. Folglich ist ein sauberes Datenmanagement auf allen Ebenen Grundvoraussetzung für eine Anwendung von MLM, um die Daten von Level 1 korrekt den Einheiten von Level 2 und diese korrekt den Einheiten von Level 3 zuzuordnen.

11 Detaillierte Darstellungen finden sich bei Luke (2004, S. 66) und König (2019b, S. 107–108).

Fazit: Unter Hinzunahme des eingangs vorgestellten fiktiven Datensatzes kann konstatiert werden, dass eine Mehrebenenmodellierung die Analyse folgender Aspekte bzw. Beantwortung folgender Fragen ermöglicht:

(1) Zunächst können alle Fragestellungen untersucht werden, die für die Analyse der Wirksamkeit der Intervention notwendig sind: Vergleich des Eingangsniveaus zum ersten Messzeitpunkt, Untersuchung der zeitlichen Entwicklung über die vier Messzeitpunkte der Experimental- und der Kontrollgruppe und die Analyse, ob die Leistungsveränderung der Experimentalgruppe signifikant anders (im Sinne von größer) als die der Kontrollgruppe ausfällt. Und es können auch nicht lineare, wie zum Beispiel quadratische oder degressive Entwicklungen modelliert werden.

(2) Darüber hinaus kann mithilfe von MLM der unterschiedliche zeitliche Abstand zwischen den vier Messzeitpunkten korrekt modelliert werden; ebenfalls führt der Einsatz von MLM zu einem besseren Umgang mit den im Datensatz vorhandenen fehlenden Werten, da diese Stichprobenelemente nicht wie bei einer rANOVA automatisch aus der Analyse ausgeschlossen werden.

Abschließend halten wir fest, dass Mehrebenenmodelle für Längsschnittdaten eine mit Blick auf die zeitliche und stochastische Struktur der Daten flexible und angemessene Analysemöglichkeit bieten, die im Rahmen von Schulsportforschung erhobene Daten gegebenenfalls angemessener abbildet als eine rANOVA.

4 | LATENTE WACHSTUMSKURVENMODELLE (LGCM)

Im Gegensatz zu den bisher vorgestellten Auswertungsverfahren arbeiten LGCM sowohl mit beobachteten als auch latenten Variablen (Faktoren)[12] und basieren auf den Grundprinzipen von Strukturgleichungsmodellen (Geiser, 2010; Kline, 2005). Sie gelten als anpassungsfähiger Ansatz zur Analyse von Veränderungen über die Zeit, wobei die Form der Veränderung über die Zeit eine zentrale Thematik darstellt. Gegenüber den etwas älteren autoregressiven und Latent-Change Modellen sind sie wesentlich flexibler (Geiser, 2010; Hertzog & Nesselroade, 1987), was folgenden Eigenschaften geschuldet ist:

- LGCM ermöglichen relativ einfache Analysen der Form der Veränderung (linear, degressiv, exponentiell …), was für sportunterrichtliche Entwicklungsprozesse relevant ist, und können diese Formen auch vergleichend prüfen. Dieser Schritt erfordert – im Gegensatz zu den bisher dargestellten Verfahren – nur relativ einfache Änderungen in der Modellsyntax (vgl. Abbildung 3 und Erklärung).
- LGCM können Veränderungen über die Zeit intra-(within person) und interindividuell (between person) modellieren und hierbei Teilstichproben vergleichen.
- LGCM erlauben auf der Basis von Korrelationsberechnungen Aussagen darüber, ob ein Ausgangswert eine Vorhersage über die Art der Veränderung erlaubt; dies wird etwa in Mplus (Muthén & Muthén, 1998–2017) automatisch ausgegeben.
- Entscheidender Vorteil ist, dass der Einsatz von LGCM den Einbau von latenten Variablen (Faktoren) in ein statistisches Modell ermöglicht.

12 Im Gegensatz zu manifesten Variablen (beobachtbare Variablen), wie zum Beispiel das Ergebnis eines sportmotorischen Tests, versteht man unter latenten Variablen nicht beobachtbare, nicht messbare und aus den beobachteten Variablen berechnete Konstrukte oder Faktoren.

- Darüber hinaus können LGCM Messfehler explizit schätzen, bei der Berechnung verschiedener Koeffizienten berücksichtigen und über die Zeit fixieren.
- Schließlich können LGCM vielfach erweitert und modifiziert werden, etwa durch eine Hinzunahme von verschiedenen Kovariaten, durch unterschiedliche Messtakte sowie durch eine Addition paralleler Wachstumskurven und deren Beziehung.

Grundprinzip von latenten Wachstumskurvenmodellen ist, dass Ausgangswert (Intercept) und Kurvensteigung (slope) als Faktoren berechnet werden, die auf einzelne Messungen (manifeste Variablen) oder Faktoren (latente Variablen) über die Zeit laden (Geiser, 2010, S. 170). Zusätzlich wird angenommen, dass alle Messungen von zufälligen Messfehlern beeinflusst werden. Dieses Grundprinzip eines LGCM 1. Ordnung, d. h. mit manifesten Variablen, ist in Abbildung 3 dargestellt.

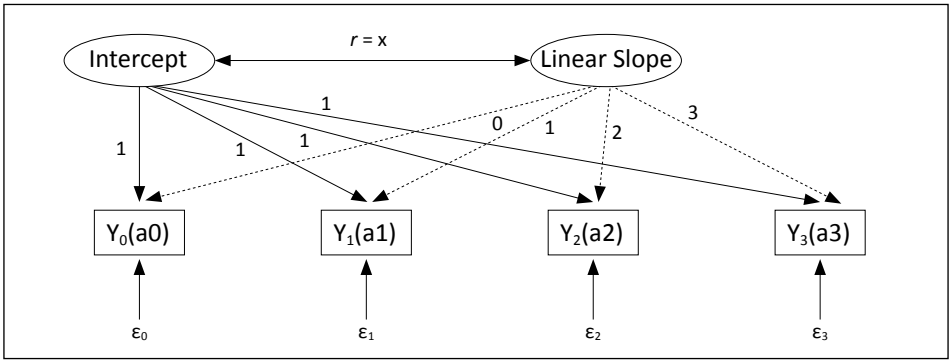

Abb. 3: Einfaches Wachstumskurvenmodell mit vier Messzeitpunkten für eine beobachtete Variable (leicht modifiziert nach Geiser, 2010, S. 170)

Das in Abbildung 3 dargestellte Basismodell zeigt die grundlegenden Konstruktionsprinzipien eines LGCM:

- Das Intercept „i" wird als Faktor berechnet, der auf jede beobachtete Variable (Messzeitpunkt) mit λ = 1 lädt; insofern geht er in jede Messung mit demselben Gewicht ein.
- Die Steigung „s" wird als Faktor berechnet, der – je nach theoretischer Wachstumsannahme – mit einer bestimmten Faktorladung auf die Messungen lädt. Abbildung 3 zeigt diesbezüglich ein angenommenes linearers Wachstum (0, 1, 2, 3). Demgegenüber würde ein quadratisches Modell die Koeffizienten 0, 1, 4 und 9 benötigen.
- Das Modell nimmt weiterhin an, dass alle Messungen von zufälligen Messfehlern beeinflusst werden; diese werden durch die Variablen $\varepsilon_{i\text{-}k}$ repräsentiert und gehen ebenfalls in die Gleichung für die abhängige Variable ein.
- Das Modell berechnet die Korrelation zwischen i und s, in Abbildung 3 als Platzhalterkoeffizient (*r* = x) angegeben. Somit kann eine Aussage bezüglich des Zusammenhangs von Ausgangswert und Entwicklungsrichtung bzw. -höhe gemacht werden; dies ist für viele sportliche Kontexte (z. B. Trainingszustand als Ausgangswert) relevant.

Dieses Basismodell kann um verschiedene Prädiktoren, wie zum Beispiel „Gruppenzugehörigkeit", um zusätzliche abhängige Variablen und um weitere Slope-Funktionen ergänzt werden. Ebenfalls besteht die Möglichkeit, die Veränderung von latenten Konstrukten, die wiederum über Variablen gemessen werden, über die Zeit zu modellieren (Geiser, 2010). Beide Varianten sind für die Sportunterrichtsforschung von Bedeutung:

- Wird in einer Studie pro Messzeitpunkt ein singulärer Wert (z. B. Ergebnis eines sportmotorischen Tests) erhoben, kann ein LGCM 1. Ordnung modelliert werden, bei dem das Intercept und die Steigung als Faktoren berechnet werden (vgl. Abbildung 3).
- Werden in einer Studie aber mehrere Messungen pro Messzeitpunkt erhoben, z. B. durch eine Testbatterie oder einen entsprechend skalierten Fragebogen, ist es möglich, ein LGCM 2. Ordnung zu berechnen, welches auf der Basis von latenten Variablen operiert. In diesem werden Faktoren 1. Ordnung aus Indikatoren berechnet und der Intercept- bzw. Slope-Faktor als Faktoren 2. Ordnung aus den Faktoren 1. Ordnung. Abbildung 4 stellt diesen Modellierungsansatz am Beispiel unserer fiktiven Studie zum Krafttraining dar.

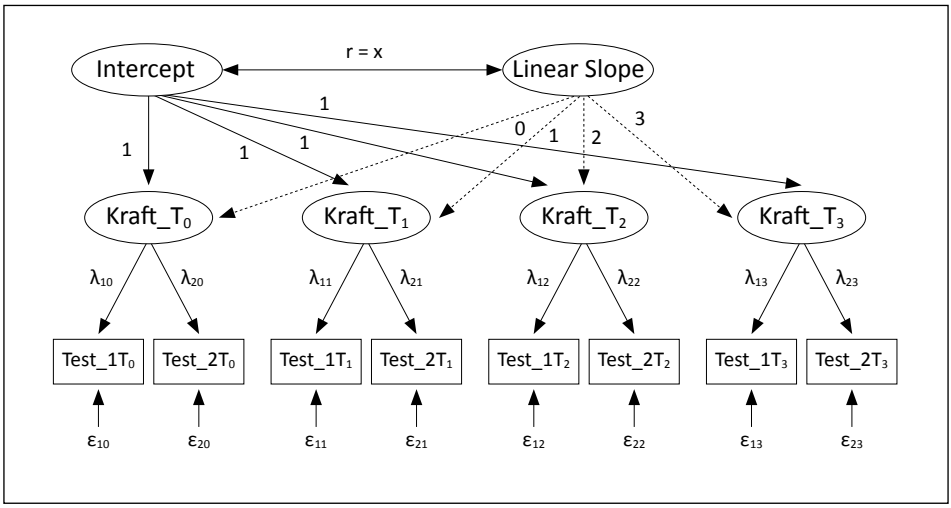

Abb. 4: Wachstumskurvenmodell 2. Ordnung (modifiziert nach Geiser, 2010, S. 189)

Zur Analyse von Wachstumskurvenmodellen empfiehlt sich die Software *Mplus* (Muthén & Muthén, 1998–2017). Mplus ist ein syntaxbasiertes Paket, welches auf einem kovarianzanalytischen Ansatz beruht. Bezüglich der Gestaltung von LGCM bietet es flexible Möglichkeiten für unterschiedliche Modellansätze, die gerade für die Sportunterrichtsforschung mit ihren manchmal schwierigen Rahmenbedingungen angemessen sind.

Fazit: Fasst man unter Hinzunahme des eingangs vorgestellten fiktiven Datensatzes zusammen, ermöglicht eine Wachstumskurvenmodellierung die Analyse folgender Aspekte bzw. die Beantwortung dieser Fragen:

(1) Ein LGCM kann sowohl die Entwicklung über einzelne sportmotorische Tests schätzen (Modell 1. Ordnung), als auch für jeden Messzeitpunkt einen Faktor „Kraft" berechnen und so die Einzelmessungen zu einer Schätzung der Veränderung über die Zeit zusammenfassen (2. Ordnung).

(2) Unabhängig von der Modellkomplexität beantwortet ein LGCM die Frage, welcher Zusammenhang zwischen dem Ausgangswert eines Individuums und der Form seiner Entwicklung besteht. Dies kann für die Sportunterrichtsforschung von besonderem Interesse sein, wenn beispielsweise Vereinsmitglieder und Nichtmitglieder hinsichtlich der Wirkung des genannten Krafttrainingsprogramms verglichen werden.

(3) Aufgrund des FIML-Algorithmus fallen fehlende Werte[13] nicht ins Gewicht, was letztendlich zu einer höheren Power und damit zu einem deutlichen Vorteil gegenüber rANOVA in der Auswertung führt.

(4) Über den Befehl „Grouping" kann ein LGCM in verschiedenste Teilstichproben aufgeteilt werden und bietet somit die Möglichkeit, die Entwicklung verschiedener Teilgruppen zu modellieren und miteinander zu vergleichen.

Abschließend kann festgehalten werden, dass LGCMs eine größere Anzahl an Möglichkeiten der Auswertung von Längsschnittdaten bereitstellen, zusätzliche interessante Koeffizienten liefern und unterschiedliche Entwicklungsformen über die Zeit anbieten. Dem stehen allerdings auch Nachteile gegenüber: Eine eventuell vorhandene Mehrebenenstruktur findet keine Berücksichtigung, eine latente Modellierung erfordert eine größere Stichprobe ($\geq$ 300) und, zumindest was *Mplus* angeht, müssen Anwender akzeptieren, dass ausschließlich mit Syntax gearbeitet wird und die Grafikfunktion nicht optimal ist.

5 | ZUSAMMENFASSUNG UND EMPFEHLUNGEN

In Bezug auf die Überlegungen zu den drei dargestellten Verfahren kann resümiert werden, dass eine Varianzanalyse mit Messwiederholung (rANOVA) ein Standardverfahren ist, welches grundsätzlich viele Möglichkeiten der Auswertung von Längsschnittdaten anbietet und relativ leicht umzusetzen ist. Demgegenüber bietet MLM eine bessere zeitliche Modellierung der Messzeitpunkte sowie die Berücksichtigung der für die Schulsportforschung typischen „nested data structure"; sie ist dafür aber komplizierter in der Anwendung und zudem sind wesentlich größere Datensätze erforderlich, was mit einem deutlich erhöhten Aufwand bei der Datenerhebung einhergeht (Kontexte auf höchster Ebene $\geq$ 30). Latente Wachstumskurvenmodelle (LGCM) haben den Vorteil, dass sie eine Analyse von latenten Konstrukten möglich machen, Messfehler berücksichtigen und hinsichtlich der Modellierung des zeitlichen Verlaufs und der Effekte als flexibel gelten; allerdings sind für LGCM 2. Ordnung mehrere Items pro Konstrukt notwendig, sodass

13 Der FIML-Algorithmus (Full Information Maximum Likelihood) ist eine Schätzmethode, in der "… missing values are not replaced or imputed, but the missing data is handled within the analysis model. The model is estimated by a full information maximum likelihood method, that way all available information is used to estimate the model. In full information maximum likelihood, the population parameters are estimated that would most likely produce the estimates from the sample data that is analysed" (Collins, Schafer & Kam, 2001).

auch sie einen erhöhten Datenerhebungsaufwand erfordern ($n \geq 300$). Zu erwähnen ist auch noch, dass bei MLM und LGCM in der Regel fehlende Werte eine weitaus geringere Rolle spielen als bei rANOVA (Allison, 2002, S. 14).

Hieraus und unter Berücksichtigung aller genannten Aspekte können folgende Empfehlungen für die Sportunterrichtsforschung abgeleitet werden:

(1) Zunächst ist nochmals in Erinnerung zu rufen, dass das beste Verfahren zur Datenauswertung nichts nützt, wenn die Qualität der erhobenen Daten unzureichend bzw. die Messungen nicht reliabel sind. Insofern müssen auch Längsschnittstudien im Kontext Schule in der Planungsphase größten Wert auf Standards in der Datenerhebung legen und es sind deshalb inhaltlich und faktoriell valide Messinstrumente zu entwickeln. Ebenfalls gilt, dass die Grundlage für Aussagen über die Wirksamkeit von Interventionen in erster Linie ein angemessenes Design ist; Quasi-Experimente sind stets nur als ausreichend und keinesfalls als optimal zu bezeichnen. Wann immer es geht, sollten deshalb experimentelle Designs umgesetzt werden (Rost, 2013), auch wenn dies im Kontext der Schulsportforschung eine große Herausforderung darstellt, da der Prozess der Randomisierung der Unterrichtsstruktur entgegensteht.

(2) Ganz entscheidend ist aus unserer Sicht, dass Studien in der Sportunterrichtsforschung zukünftig mit mehr als zwei Messzeitpunkten durchgeführt werden. Uns ist bewusst, dass forschungsökologische Gründe keine „grenzenlose" Erweiterung der Anzahl an Messungen erlauben, dennoch ist eine Studie mit zwei Datenerhebungen nicht robust belastbar (Singer & Willett, 2003, S. 10). Außerdem lassen zwei Messzeitpunkte keinerlei Rückschlüsse auf den Verlauf möglicher Veränderungen zu, was aber gerade für Lern- und Trainingsprozesse eine wichtige Information sein kann – etwa um Mindestzeiträume für eine Intervention zu kennen und einzuplanen. Selbstverständlich ist diese Aussage für klassische Experimente obsolet.

(3) Zu ergänzen ist, dass der Messzeittakt (Abstand zwischen zwei Messungen) theoriegeleitet festgelegt werden und sich eng an der Struktur sowie der zeitlichen Stabilität des Forschungsgegenstandes (z. B. motorische oder kognitive Lernprozesse) orientieren muss. Auch wenn schulische Rahmenbedingungen (z. B. Ferieneinschnitte, Kursangebote) häufig die Forschungspraxis dominieren, dürfen Veränderungen über Zeit nicht „theorielos" beobachtet werden, sondern müssen – wenn irgendwie möglich – solche theoretischen Gegebenheiten bei der Modellierung des Messzeittakts berücksichtigen.

(4) Weiterhin sollte bei Längsschnittstudien in der Schulsportforschung bei der Planung bzw. der Kalkulation der Stichprobe verstärkt auf statistische Zusammenhänge zwischen zu erwartendem Effekt, α- und β-Niveau sowie Testpower geachtet werden. Eine Ad-hoc-Stichprobenziehung, die sich lediglich an den Möglichkeiten vor Ort orientiert[14], reduziert die Belastbarkeit der Ergebnisse bzw. kann zu Verzerrungen führen; inferenzstatistische Verfahren setzen echte Zufallsstichproben voraus. Auch ist bei der Größe der Stichprobe stets an fehlende Werte und deren

14 In der Praxis von Schulsportforschung ist die Realisierung echter Zufallsstichproben, was auch aus methodischer Sicht äußerst wünschenswert wäre, i. d. R. aus Gründen der Freiwilligkeit der Teilnahme der Schulen bzw. der Schüler*innen kaum möglich.

Konsequenzen zu denken. Diese Überlegungen zur Stichprobenkalkulation gelten in besonderem Maße für Mehrebenenmodelle, die die Gegebenheiten der Schulsportforschung („nested data“) im Gegensatz zu Einebenenverfahren berücksichtigen. Eine entscheidende Empfehlung (u. a. Hox, 2010; König, 2019b; Maas & Hox, 2005; McNeish & Stapleton, 2016; Snijders & Bosker, 2012) ist, die Stichprobengröße auf der höchsten Ebene als die limitierende Größe zu betrachten und hier wenigstens 30, eher 50 Kontextgruppen zu integrieren. Als Konsequenz muss auf den Ebenen darunter nicht jeder einzelne Fall getestet werden.

(5) Liegen nach sorgfältiger Planung und Erhebung die Daten vor, ist das Auswertungsverfahren entsprechend der Datenstruktur zu wählen; dies bedeutet, dass es keine Automatismen geben darf, sondern Gegebenheiten, wie etwa Schachtelung, Datenkomplexität oder Auswertungsinteresse, die Auswertung leiten. Folglich müssen Daten zunächst hinsichtlich entsprechender Merkmale statistisch geprüft werden (z. B. durch Berechnung des Intraklassenkoeffizienten ICC), bevor ein komplexeres Verfahren unreflektiert zum Einsatz kommt (Magnaguagno et al., 2016, S. 60). Schließlich ist zu jedem Zeitpunkt der Studie abzuwägen, ob ein methodologisch oder methodisch notwendiger Mehraufwand (z. B. Berechnung von latenten Konstrukten zu mehreren Messzeitpunkten) bezüglich der Robustheit, aber auch der Aussagekraft der Ergebnisse gerechtfertigt ist; auch bei Längsschnittstudien in der Sportunterrichtsforschung gilt das Gebot der Sparsamkeit.

Werden alle Argumente abschließend subsumiert, wird deutlich, dass hochwertige Längsschnittstudien in der Schulsportforschung echte Herausforderungen darstellen; wer aber von Anfang an systematisch und sorgfältig die genannten Empfehlungen in Planung und Realisierung berücksichtigt, kann sichergehen, dass dadurch viele Fußangeln und Fallstricke umgangen werden.

LITERATUR

Allison, P. D. (2002). *Missing Data* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07–136). Sage.

Aschebrock, H., & Stibbe, G. (Hrsg.). (2018). *Schulsportforschung. Wissenschaftstheoretische und methodologische Reflexionen*. Waxmann.

Bähr, I., Bund, A., Gerlach, E., & Sygusch, R. (2011). Evaluationsforschung im Sportunterricht. In E. Balz, M. Bräutigam, W.-D. Miethling & P. Wolters (Hrsg.), *Empirie des Schulsports* (S. 44–63). Meyer & Meyer.

Balz, E., Bräutigam, M., Miethling, W.-D., & Wolters, P. (2011). *Empirie des Schulsports*. Meyer & Meyer.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation* (4., überarb. Aufl.). Springer.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Springer.

Brandl-Bredenbeck, H. P., & Stefanie, M. (Hrsg.). (2009). *Schulen in Bewegung – Schulsport in Bewegung. Jahrestagung der dvs-Sektion Sportpädagogik vom 22.–24.05.2008 in Köln*. Czwalina.

Bräutigam, M. (2008). Schulsportforschung – Skizze eines Forschungsprogramms. In Dortmunder Zentrum für Schulsportforschung (Hrsg.). *Schulsportforschung. Grundlagen, Perspektiven und Anregungen* (S. 14–50). Meyer & Meyer.

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

Conzelmann, A., Hänsel, F., & Höner, O. (2013). Individuum und Handeln – Sportpsychologie. In A. Güllich & M. Krüger (Hrsg.), *Sport. Das Lehrbuch für das Sportstudium* (S. 269–336). Springer.

Cronbach, L., Deken, J., & Webb, N. (1976*). Research on Classrooms and Schools: Formulation of Questions, Design and Analysis*. Occasional papers of the Stanford evaluation consortium. Stanford Evaluation.

Diekmann, A. (2011). *Empirische Sozialforschung. Grundlagen – Methoden – Anwendungen* (19., vollständig überarb. und erw. Aufl.). Rowohlt.

Eid, M., Gollwitzer, M., & Schmitt, M. (2013). *Statistik und Forschungsmethoden* (3., korrig. Aufl.). Beltz.

Field, A. (2014). *Discovering Statistics Using IBM SPSS Statistics* (4. Aufl.). Sage.

Fitzmaurice, G. M., Laird, N. M., & Ward, J. H. (2011). *Applied Longitudinal Analysis* (2. Aufl.). Wiley.

Friedrich, G. (2000). Schulsportforschung – Zur Konzeption eines ausbaubedürftigen Bereichs der Sportwissenschaft. *dvs-Informationen*, *15*(1), 7–11.

Friedrich, G. (2002). (Hrsg.). *Sportpädagogische Forschung. Konzepte – Ergebnisse – Perspektiven* (S. 274–280). Czwalina.

Garson, G. D. (2013). Fundamentals of Hierarchical Linear and Multilevel Modeling. In G. D. Garson (Eds), *Hierarchical Linear Modeling. Guide and Applications* (S. 3–25). Sage.

Geiser, C. (2010). *Datenanalyse mit Mplus. Eine anwendungsorientierte Einführung* (2., durchgesehene Aufl.). VS-Verlag.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* Oxford University Press.

Helmke, A. (2010). *Unterrichtsqualität. Erfassen – Bewerten – Verbessern*. Kallmeyer.

Hemphill, M. A., Richards K. A. R., Templin, T. J., &. Blankenship, B. T. (2012). A Content Analysis of Qualitative Research in the Journal of Teaching in Physical Education from 1998 to 2008. *Journal of Teaching in Physical Education, 31*, 279–287.

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the state-trait distinction for the structural modelling of developmental change. *Child Development, 58*, 93–109.

Hirtz, P., & Forschungszirkel „N. A. Bernstein“ (2007). *Phänomene der motorischen Entwicklung des Menschen*. Hofmann.

Hox, J. J. (2010). *Multilevel Analysis – Techniques and Applications*. Routledge.

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Johnson, R. B., & Christensen, L. (2014). *Educational Research: Quantitative, Qualitative, and Mixed Approaches* (5. Aufl.). Sage.

Kalaja, S., Jaakkola, T., Liukkonen, J., & Digelidis, N. (2012). Development of junior high school students' fundamental movement skills and physical activity in a naturalistic physical education setting, *Physical Education and Sport Pedagogy, 17*(4), 411–428. https://doi.org/10.1080/17408989.2011.603124

Kalnbach, T. (2019). *Implizites Training im Sportunterricht. Eine Studie zum Thema Trainieren im Schulsport*. Logos.

Keller, F. (2004). Analyse von Längsschnittdaten: Auswertungsmöglichkeiten mit hierarchisch linearen Modellen. *Zeitschrift für Klinische Psychologie und Psychotherapie, 32*(1), 51–61.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed.). The Guilford Press.

König, S. (2002). Sportunterrichtsforschung: Unverzichtbarer Bestandteil der Sportlehrerausbildung. In G. Friedrich (Hrsg.), *Sportpädagogische Forschung. Konzepte – Ergebnisse – Perspektiven* (S. 274–280). Czwalina.

König, S. (2011). *Körperliche Förderung im Schulsport. Theoretische Ansätze, empirische Studien und praktische Konzepte zur Unterrichtsentwicklung*. Logos.

König, S. (2016). Koordinationstraining im Schulsport. In G. Thienes & M. Baschta (Hrsg.), *Training im Schulsport* (S. 139–158). Hofmann.

König, S. (2019a). Evaluating Fitness Training in Physical Education – A Quantitative Dominated Crossover Mixed Methods Multilevel Study. *International Journal of Multiple Research Approaches, 11*(1), 45–60.

König, S. (2019b). Mehrebenenanalysen: Unterrichtsforschung im Fach Sport zur Effektivität von Trainingsprozessen. In G. Lang-Wojtasik & S. König (Hrsg.), *Die Vielfalt methodischer Zugänge in der Unterrichtsforschung. Beispiele aus Fachdidaktik und Allgemeiner Didaktik. Weingartner Dialog über Forschung 3* (S. 101–122). Klemm + Oelschläger.

Lindel, M. (2018). *Methoden der Sportunterrichtsforschung im Vergleich – eine Untersuchung am Beispiel von Sound-Karate*. Logos.

Lüdtke, O., Trautwein, U., Schnyder, I., & Niggli, A. (2007). Simultane Ebenen auf Schüler- und Klassenebene. Eine Demonstration der konfirmatorischen Mehrebenen-Faktorenanalyse zur Analyse von Schülerwahrnehmungen am Beispiel der hausaufgabenvergabe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39*(1), 1–11.

Luke, D. A. (2004). *Multilevel Modeling* (Sage University Paper Series on Quantitative Applications in Social Sciences, (S. 07–143). Sage.

Lutz, M. (2018). *Effekte von Textproduktion auf das Bewegungslernen*. Logos.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(2), 85–91.

Magnaguagno, L., Schmidt, M., Valkanover, S., Sygusch, R., & Conzelmann, A. (2016). Programm- und Outputevaluation einer Intervention zur Förderung des sozialen Selbstkonzepts im Sportunterricht. *Zeitschrift für Sportpsychologie*, *23*(2), 56–65.

Maxwell, S. C., & Delaney, H. D. (2004). *Designing Experiments and Analyzing Data. A Model Comparison Perspective* (2. Aufl.). Taylor & Francis.

McNeish, D., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two Level Model Estimates: A Review and Illustration. *Educational Psychology Review, 28*(2), 295–314.

Meinel, K., & Schnabel, G. (Hrsg.). (2007). *Bewegungslehre – Sportmotorik* (11., überarb. und erw. Aufl.). Meyer & Meyer.

Memmert, D., & König, S. (2007). Teaching Games in Elementary Schools. *International Journal of Physical Education*, *44*, 54–66.

Muthén, L., & Muthén, B. (1998–2017). *Mplus User's Guide* (8. Aufl.). Muthén & Muthén.

Novak, D., & Bernstein, E. (2015). A Review of Research on Physical Education Teachers and Coach Education (2013–2014). *International Journal of Physical Education LII*(3), 2–9.

Onwuegbuzie, A., & Hitchcock, J. (2015). Advanced Mixed Analysis Approaches. In S. N. Hesse-Biber & R. B. Johnson (Eds.), *Oxford Handbook of Multimethod and Mixed Methods Research Inquiry* (S. 275–295). Oxford University Press.

Oesterhelt, V. (2011). Schneesport in der Schule im Rahmen des erziehenden Sportunterrichts – Evaluation eines didaktisch-methodischen Unterrichtskonzepts. In F. Borkenhagen, S. Hafner, R. Heim & P. Neumann (Hrsg.), *Kinder- und Jugendsport zwischen Gegenwarts- und Zukunftsorientierung* (S. 59). Czwalina.

Prohl, R. (2010). *Grundriss der Sportpädagogik* (3., korr. Aufl.). Limpert.

Prohl, R. (2012). Der Doppelauftrag des Erziehenden Sportunterrichts. In R. Prohl & V. Scheid (Hrsg.), *Sportdidaktik. Grundlagen – Vermittlungsformen – Bewegungsfelder* (S. 70–91). Limpert.

Raithel, J. (2008). *Quantitative Forschung. Ein Praxiskurs* (2. Aufl.). VS Verlag für Sozialwissenschaften.

Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2014). Quantitative Methoden 2. Einführung in die Statistik für Psychologen und Sozialwissenschaftler (4. Aufl., 2 Bände). Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2. Aufl.). Sage.

Rost, D. (2013). *Interpretation und Bewertung pädagogisch-psychologischer Studien: Eine Einführung* (3. Aufl.). Julius Klinkhardt.

Schendera, C. (2008). *Regressionsanalysen mit SPSS*. Oldenbourg Verlag.

Schiemann, S., & Pargätzi, J. (2016). Beweglichkeitstraining im Schulsport. In G. Thienes & M. Baschta (Hrsg.), *Training im Schulsport* (S. 116–138). Hofmann.

Schmid, J., Haible, S., & Sudeck, G. (2020). Patterns of physical activity-related health competence: stability over time and associations with subjective health indicators. *German Journal of Exercise and Sport Research.* https://doi.org/10.1007/s12662-020-00650-1

Schnell, R., Hill, P., & Esser, E. (2011). *Methoden der empirischen Sozialforschung* (7. Aufl.). R. Oldenbourg.

Sedlmeier, P., & Renkewitz, R. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler. Ein Lehrbuch für Psychologen und Sozialwissenschaftler* (2. Aufl.). Pearson.

Silverman, S., & Skonie, R. (1997). Research on Teaching in Physical Education: An Analysis of Published Research. *Journal of Teaching in Physical Education*, *16*, 300–311.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis*. *Modeling Change and Event Occurrence.* Oxford University Press.

Snijders, T., & Bosker, R. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2. Aufl.). Sage.

Thiele, J. (2008). Formen der Erkenntnisgewinnung in der Schulsportforschung – Methodologie und Methoden. In Dortmunder Zentrum für Schulsportforschung (Hrsg.), *Schulsportforschung. Grundlagen, Perspektiven und Anregungen* (S. 51–72). Meyer & Meyer.

Thienes, G. (2016). Schnelligkeitstraining im Schulsport. In G. Thienes & M. Baschta (Hrsg.), *Training im Schulsport* (S. 178–197). Hofmann.

Töpfer, C., Bähr, I., König, S., Reuker, S., & Sygusch, R. (2020, preprint). Interventionsstudien im Sportunterricht. In E. Balz, C. Krieger, W.-D. Miethling & P. Wolters (Hrsg.), *Empirie des Schulsports*. Meyer & Meyer.

Twisk, J. (2010). Experimental Methods. In H. Haag (Ed.), *Research Methodology for Sport and Exercise Science. A Comprehensive Introduction for Study and Research* (S. 194–222). Logos.

Walker, E., & Nowacki, A. S. (2011). Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine*, *26*(2), 192–196.

Wartenberg, J., Borchert, T., & Brand, R. (2014). A longitudinal assessment of adolescent student-athletes' school performance. *sportwissenschaft, 44*(2), 78–85.

Weißeno, G. (2019). Strukturgleichungsmodelle: Unterrichtsforschung im Fach Politik zum politischen Wissen über Demokratie. In Lang-Wojtasik, G., & König, S. (Hrsg.), *Die Vielfalt methodischer Zugänge in der Unterrichtsforschung. Beispiele aus Fachdidaktik und Allgemeiner Didaktik* (S. 85–100). Klemm + Oelschläger.

Wirszing, D. (2015). *Die motorische Entwicklung von Grundschulkindern. Eine längsschnittliche Mehrebenenanalyse von sozioökologischen, soziodemographischen und schulischen Einflussfaktoren*. Czwalina.

Wolters, P. (2011). Unterrichtsforschung. In E. Balz, M. Bräutigam, W. D. Miethling & P. Wolters (Hrsg.), *Empirie des Schulsports* (S. 19–43). Meyer & Meyer.