Zur Interpretation von Messfehlern aus Sicht der Erziehungswissenschaft

Michael Schurig & Daniel Kasper

1 Einführung

In Abgrenzung zur akademischen Psychologie interessiert sich die erziehungswissenschaftliche Forschung nicht zentral für inter- oder intraindividuelle Prozesse und in Abgrenzung zur Sozialwissenschaft nicht für gesellschaftliche systemimmanente Phänomene, sondern vor allem für die Interaktion zwischen Individuum und Umwelt, insbesondere für den Einfluss pädagogischer Prozesse unter explizitem Einbezug der darin verankerten Reflexivität gegenüber den im Bildungsbegriff indizierten Aufgabenstellungen (Zedler/Döbert 2010, S. 40). Also werden komplexe Gefüge des Ganzen pädagogischer Prozesse betrachtet, was im Sinne der empirischen Bildungsforschung pragmatisch beschrieben werden muss (ebd.; Edelmann/Schmidt/Tippelt 2012, S. 60). Für statistische Analysen bedeutet diese Pragmatik eine Reduktion der Realität in eine formale Struktur, welche mit der Annahme verknüpft ist, dass Messfehler existieren und deren Angemessenheit über die Hauptgütekriterien von Tests zu belegen ist. Die Berücksichtigung von Messfehlern wird überdisziplinär als notwendig erachtet, wenn in Analysen ein idiosynkratischer Rahmen verlassen werden soll. Die Annahmen zu den Eigenschaften der Fehler sind allerdings an fachspezifische Interpretationskulturen geknüpft. So müssen, wenn ganze pädagogische Prozesse betrachtet werden sollen, Annahmen zur Zufälligkeit und Zusammenhangslosigkeit von Fehlern häufig zurückgewiesen werden. Insbesondere angesichts der Komplexitätsgrade von Prozessen zwischen Individuen, pädagogischen Akteuren, Systemen und Lebenswelten drängt es sich auf, dass Fehler stärker interpretiert und analysiert werden, um möglichst valide Evidenz (Newton/Shaw 2013) für die Annahme von Wahrheitsaussagen zu erzeugen.

In dem vorliegenden Beitrag soll der theoretische Rahmen zur Interpretation von Messfehlern aus Sicht der Erziehungswissenschaft dargestellt werden. Darauf aufbauend wird argumentiert, dass Fehler intensiver analysiert werden sollten und dass deren Bewertung kein rein statistischer, sondern ein substanzwissenschaftlicher Prozess ist.

 https://doi.org/10.3224/ezw.v29i1.06

2 Messfehler und deren Bedeutung

Prominente Theorien der Psychometrie gehen davon aus, dass Messungen nicht fehlerfrei Erfolgen können (Lord/Novick 1968). Erst die Annahme der Existenz von Messfehlern ermöglicht es, quantitative Analysen durchzuführen, ohne Inkonsistenzen aufzuwerfen (Kane 2011). Wenn beispielsweise eine zeitstabile Eigenschaft zu zwei verschiedenen Zeitpunkten gemessen wird, die Messergebnisse aber variieren, gibt es verschiedene Ansätze, dies zu erklären. Die Variabilität kann konzeptionell zufälligen Messfehlern zugeschrieben werden, als mangelhafte Sensitivität des Instrumentes oder der Messung, gegebenenfalls in Abhängigkeit zur gemessenen Eigenschaft (Embretson 1996, S. 342), oder als Funktion von Drittvariablen verstanden werden (Kane 2011). Diese Annahmen sind dabei an das den Analysen zugrunde gelegte testtheoretische Modell geknüpft. So wird beispielsweise bei der Vorhersage einer Variable im allgemeinen linearen Modell angenommen, dass sich der Fehleranteil einer Vorhersage aus dem Spezifikationsfehler, dem Messfehler und einem Zufallsanteil zusammensetzt (Werner 1997). Allgemeiner trennte Nunnally (1978) den Fehleranteil eines Messergebnisses in systematische und zufällige Fehler. Systematische Fehler umfassen konsistente Unterschiede zwischen Gruppen innerhalb der Population, welche nicht mit dem untersuchten Konstrukt zusammenhängen. Zufällige Fehler werden hingegen als darüber hinausgehende Abweichung zwischen dem Messwert und dem wahren Wert auf der Individualebene interpretiert. Ein systematischer Fehler könnte beispielhaft auftreten, wenn zur Lösung einer Mathematikaufgabe eine hohe Lesefähigkeit benötigt wird. Ein zufälliger Fehler könnte beispielsweise im zufälligen Raten einer Aufgabenlösung begründet liegen.

In den Axiomen der klassischen Testtheorie wird zur Lösung der Gleichungssysteme angenommen, dass der Messfehler eine Komponente einer Messung ist, welche keine Kovarianz mit der Messgröße aufweist und unabhängig von anderen Messungen ist (z. B. Lienert/Raatz 1998). Diese Sichtweise ist mathematisch relativ einfach und praktisch bewährt, aber nicht unproblematisch. So würden beispielsweise unberücksichtigte systematische Fehler den individuellen wahren Werten zugerechnet werden. Die Reliabilität, also das Verhältnis der ‚wahren' Varianz zu der beobachteten Varianz der Stichprobe, stiege also bei einem starken Einfluss systematischer Fehler. Distinkte Fehler in den Prädiktoren bleiben ebenso unberücksichtigt, können aber in Form von unterschiedlich trennscharfen Zusammenhängen der Prädiktoren zu der interessierenden Eigenschaft (Heteroskedastizität) oder in Form von Zusammenhängen zu derselben Eigenschaft zu einem früheren Zeitpunkt (Autokorrelation) auftreten. Für Analysen komplexer Prozesse sind zufällige und systematische Fehleranteile sowie Fehleranteile in Prädiktoren und in der Vorhersage zu trennen. Durch generalisierte Anwendungen, wie

Modelle latenter Variablen, wird diese Forderung aufgegriffen; so können differenzierte Fehleranteile, Residuale, geschätzt werden.

Doch entsprechend der repräsentativen Messtheorie, wonach eine Messung eine numerisch strukturerhaltende Operationalisierung der Eigenschaft eines Subjekts (Stevens 1946) darstellt, deutet nichts darauf hin, dass ein Wert messfehlerbehaftet sein könnte (Kane 2011). Die Annahme des Vorhandenseins zufälliger und systematischer Fehler stellt eine inhaltliche Setzung da (ebd.). Diese füllt die inhaltslosen Fehler mit Sinn und macht sie ihrerseits zu nicht in den Daten direkt beobachtbaren, also latenten (vgl. Bollen 2002) Variablen.

3 Kontrolle für systematische Messfehler

Für die Kontrolle von Drittvariablen, welche systematische Fehler bedingen können, stellt das randomisierte Experiment den Gold-Standard da (z. B. Cronbach 1982). Da angenommen wird, dass Drittvariablen zufällig auf Experimental- und Kontrollgruppe gleichverteilt werden können, wäre ein Effekt gleich der Differenz der Erwartungswerte in der Experimental- und der Kontrollgruppe, da nur die unabhängige Variable zwischen den Gruppen variiert. Die Umsetzung ist aber in der Bildungsforschung aus praktischen Gründen häufig nicht möglich (Bromme/Prenzel/Jäger 2014, S. 14). Dies kann zum Beispiel im Zeitaufwand, in ethischen Gründen bei der Wahl der Zuweisungsmechanik oder in der Unkontrollierbarkeit der Forschungsumgebung begründet sein. So kann, in Abgrenzung zu idealtypischen experimentellen Studien, schon aufgrund der Stichprobenziehung (häufig möglichst intakte Lerngemeinschaften), nicht angenommen werden, dass systematische Fehler aufgrund von Randomisierungen ausgeschlossen werden können, da Effekte auf Klassen- oder Schulebene vorliegen können. Der Auftrag der erziehungswissenschaftlichen Bildungsforschung ist außerdem die Bereitstellung handlungsleitender Informationen. Dafür sollten die gewonnenen Erkenntnisse möglichst generalisierbar sein, was für experimentelle Studien nur eingeschränkt gilt. Eine experimentelle Studie ist auch immer eine atypische Studie, welche für die in der Bildungsforschung besonders relevante Evaluations- und Interventionsforschung ungeeignet sein kann (Cronbach 1982).

Wenn nur bedingt eine Kontrolle für forschungsrelevante Drittvariablen vorgenommen werden kann, müssen beobachtete Werte als Schätzungen generellerer Eigenschaften über unbeobachtete Drittvariablen hinweg betrachtet werden und jedwede Variabilität, die nicht auf die generelle Eigenschaft zurückgeführt werden kann, wird als Fehler betrachtet (Kane 2011). Es ist aber üblich, die Fehler(ko)varianzen explizit zu modellieren, deren Größe zu bestimmen und zur Bewertung der Prädiktoren und der Modellanpassung zu nutzen.

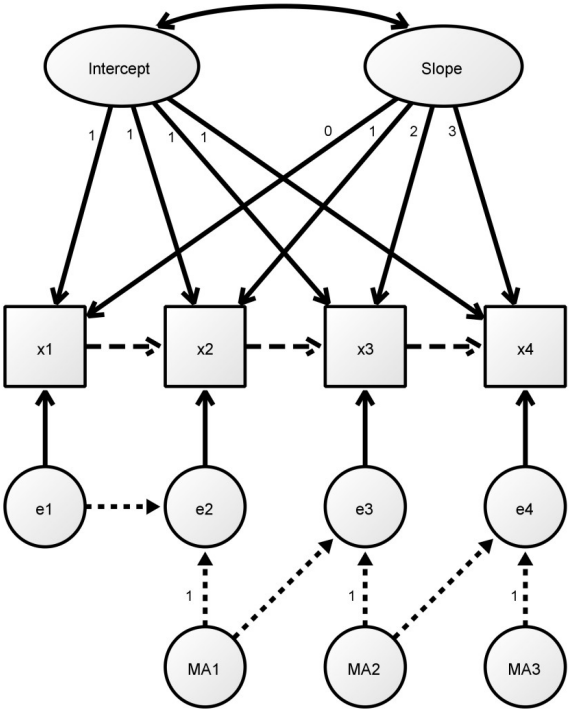
4 Modellierung von Messfehlern

Flexible quantitative Auswertungsverfahren ermöglichen die Verarbeitung verschiedener Annahmen gegenüber den Fehlern, wie die Lockerung der Restriktion, dass Fehlervarianzen über verschiedene Indikatoren theoretischer Konstrukte nicht variieren dürfen, wie etwa bei Operationalisierungen über Summen- oder Mittelwertbildungen (Steyer/Eid 2001). Zudem kann beobachtet werden, dass die Berücksichtigung von Clusterstrukturen innerhalb von Stichproben, beispielsweise in Mehrebenenmodellen oder durch die Korrektur von Standardfehlern, in der bildungswissenschaftlichen Forschung inzwischen fest verankert sind (Windzio/Teltemann 2013). Dies erlaubt z. B. eine Kontrolle oder Korrektur für bekannte Gruppenzugehörigkeiten. Besondere Relevanz haben Fehlerstrukturen auch in längsschnittlichen Analysen. So würde der Einfluss der Umwelt auf eine Entwicklung entweder durch eine Drittvariable, also einen eigenen Faktor, repräsentiert oder in die Fehler einfließen. Wird kein Faktor für die Umwelt postuliert, dann müssten Fehler vorheriger (Mess)Zeitpunkte auf die abhängige Variable zum gegenwertigen Zeitpunkt wirken und gegebenenfalls auch korrelieren. Analytisch könnten Latent Growth Curve Models (LGCM; Meredith/Tisak 1990) zur messtheoretisch fundierten Abbildung einer angenommenen Entwicklung herangezogen werden.

Die Erziehungswissenschaft würde aber in natürlicher Weise davon ausgehen, dass beispielsweise eine Veränderung über die Zeit auch auf eine Veränderung der Umwelt im selben Zeitraum zurückzuführen ist. So könnte eine Schule beispielsweise ihr pädagogisches oder eine Lehrkraft ihr didaktisches Konzept ändern. Dabei kann die Struktur der Fehler jenseits einfacher korrelativer Zusammenhänge mittels autoregressiver (AR) und „Moving Average“ (MA) Komponenten aufgenommen werden (Werner 2005).

Die autokorrelativen Zusammenhänge können dabei in unterschiedlicher Weise verarbeitet werden, z. B. in erster oder zweiter Ordnung und allein unter Berücksichtigung von AR oder MA Prozessen oder inklusive beider. Diese Unterschiede können einen gewichtigen Einfluss auf die Schätzungen der Effekte im Modell haben (Kwok/West/Green 2007). In der Abbildung 1 ist ein LGCM inklusive AR- (gestrichelte Linien) und MA-Prozesse (gepunktete Linien) dargestellt.

Abbildung 1: Lineares Wachstumskurvenmodell mit ARMA Struktur erster Ordnung



Quelle: Sivo/Fan/Witta 2005

In dem Modell könnte x_t (t =1,…,4) etwa die mathematische Kompetenz zum Zeitpunkt t symbolisieren, e_t würde den Fehler zum Zeitpunkt t darstellen und Intercept bzw. Slope bilden systematische inter- bzw. intraindividuelle Unterschiede in der mathematischen Kompetenz ab. So wird beispielsweise angenommen, dass die mathematische Kompetenz zum Zeitpunkt t=1 einen Effekt auf die mathematische Kompetenz zum Zeitpunkt t=2 hat (AR) und der Fehler zum Zeitpunkt t=1 einen Effekt auf den Fehler zum Zeitpunkt t=2 hat (MA). Der letztgenannte Effekt könnte beispielsweise als Folge einer systematisch veränderten pädagogischen Praxis interpretiert werden, die gleichwohl einen Einfluss auf die Kompetenzentwicklung der Schülerinnen und Schüler hat.

Unter der Berücksichtigung des Ganzen pädagogischer Prozesse, also beispielsweise unter explizitem Einbezug von Kontextbedingungen des Lernens, welche aber ihrerseits nicht im direkten Forschungsfokus stehen, ist eine ARMA Struktur nicht nur gut zu interpretieren, ein solches Modell wäre geradezu zu erwarten und analytisch gut umsetzbar. LGCM mit ARMA-Strukturen wurden bereits in zahlreichen Arbeiten der pädagogischen Psychologie angewendet (Grimm/Widaman 2010; Marsh/Grayson 1994; Sivo/Willson 2000). Die Interpretation der ARMA-Strukturen im Sinne der Erziehungswissenschaft ist unseres Wissens aber neu und würde theoretischen Annahmen zu ganzheitlichen pädagogischen Wirkprozessen Rechnung tragen. Auf diese Weise kann der Sichtvorteil der Erziehungswissenschaft gegenüber anderen Disziplinen (Zedler/Döbert 2010), also der explizite Einbezug kontextueller und normativer Rahmenbedingungen, auch stärkeren Eingang in die formalen Analyseprozesse erlangen.

5 Diskussion

Das Ziel jeder Methode muss eine datengestützte Fehlerreduktion, also eine „Wahrheitsfindung“ sein, wobei die Daten als Gegner der eigenen Annahmen begriffen werden (Fend 2009, S. 27). Wie Kane (2011) eindrucksvoll darstellen kann, sind Fehler aber keine empirischen Fakten, die per se an den Daten hängen, sondern Resultate von interpretativen Prozessen und Reflexionen, welche durch den Zufall oder Eigenschaften der Messung, des Konstrukts sowie der Untersuchungseinheiten determiniert sein können. Fehler existieren nicht, bis wir sie erschaffen (ebd.). Sie werden erschaffen, um Inkongruenzen zwischen unseren mathematischen Modellen und der Realität zu erklären. Fehler werden durch fachspezifische theoretische Annahmen mit Sinn erfüllt und durch Varianzanteile in testtheoretisch fundierten Modellen repräsentiert.

Dabei gebietet der Untersuchungsgegenstand des Ganzen pädagogischer Prozesse vor dem Hintergrund eines reflexiven und handlungsleitenden Forschungsverständnisses, dass Fehleranteile mit analysiert werden. Insbesondere, da Strategien zur Minimierung systematischer Fehler häufig nicht oder nur mit großem Aufwand, z. B. durch die Parallelisierung von Schulklassen, angewendet werden können. Durch die Integration von inhaltlich begründeten Fehlern in die Analyseprozesse kann es gelingen, mathematische Modelle realitätsgetreuer zu gestalten. Die pädagogische Realität ist komplex; unsere Modelle müssen aber einfach bleiben, um bearbeitbar zu sein. Gleichzeitig stellen unrealistische Modellrestriktionen eine Gefahr für die Validität wissenschaftlicher Wahrheitsaussagen dar. Basierend auf dem zur Verfügung stehenden flexiblen quantitativen Modellrepertoire, besteht keine Notwendigkeit generell und ritualhaft Annahmen zur Invarianz, Zufälligkeit und Zusammenhangslosigkeit von Fehlern vorzunehmen. Die Entscheidungen darüber

können sich hingegen inhaltlich und aus der Relevanz und dem Gewicht der Fehler ergeben. Fehler müssen im Verhältnis zur Toleranz der Analysen klein sein. Wenn eine Fehlerquelle im Verhältnis zu maßgeblichen Fehlerquellen in einem akzeptablen Modell klein ist, kann diese ignoriert werden, damit die Modelle nicht überkomplex werden (Kane 2011).

Diese Relativität ist vielfältig kritisiert worden (Skrondal/Rabe-Hesketh 2004, S. 1), doch letztlich obliegen Annahmen über die Angemessenheit einer Operationalisierung subjektiv den Forschenden (Kline 2011, S. 191). Diese Relativität setzt Kenntnisse über die Funktionsweisen von Modellgüteindizes und besonders den substantiellen Gegenstandsbereich und dessen kontextueller und normativer Bedingungen voraus, denn insoweit über einen Sachverhalt ein relevanter Wissensbestand fehlt oder eine Repräsentation nicht verfügbar ist, kann dieser in den Hypothesen nicht auftauchen (Kelle 2008, S. 34). Die Urteile über die Fehlertoleranzen sind qualitativer Natur und basieren auf domänenspezifischem Wissen vor dem Hintergrund quantifizierter Unsicherheit (Schurig 2017, S. 86). Und auch wenn keine normativen allgemeingültigen Schwellen für die Bewertung einer akzeptablen Größe von Fehleranteilen existieren, liegt hier keine Beliebigkeit, sondern ein Rückbezug auf die Hauptgütekriterien des Testens und Messens vor. „An angemessenen wissenschaftlichen Standards, die auch disziplinübergreifend anzuwenden sind, führt kein Weg vorbei“ (Schwippert 2016, S. 36).

Michael Schurig, Dr. phil., ist Wissenschaftlicher Mitarbeiter an der Fakultät für Erziehungswissenschaft, Psychologie und Soziologie, Institut für Schulentwicklungsforschung an der TU Dortmund.

Daniel Kasper, Dr. phil., ist Wissenschaftlicher Mitarbeiter an der Fakultät für Erziehungswissenschaft, Allgemeine, Interkulturelle und International Vergleichende Erziehungswissenschaft an der Universität Hamburg.

Literatur

Bollen, Kenneth A. (2002): Latent variables in psychology and the social sciences. In: Annual Review of Psychology 53, S. 605-634. http://dx.doi.org/10.1146/annurev.psych.53.100901.135239.

Bromme, Rainer/Prenzel, Manfred/Jäger, Michael (2014): Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. In: Zeitschrift für Erziehungswissenschaft 17, 4, S. 3-54. http://dx.doi.org/10.1007/s11618-014-0514-5.

Cronbach, Lee J. (1982): Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass.

Edelmann, Doris/Schmidt, Joel/Tippelt, Rudolf (2013): Einführung in die Bildungsforschung. Stuttgart: Kohlhammer.

Embretson, Susan E. (1996): The new rules of measurement. In: Psychological Assessment 8, 4, S. 341-349. http://dx.doi.org/10.1037/1040-3590.8.4.341.

Fend, Helmut (2009): Bildungsforschung von 1965 bis 2008. In: Wischer, B./Tillmann, K.-J. (Hrsg.): Erziehungswissenschaft auf dem Prüfstand. Schulbezogene Forschung und Theoriebildung von 1970 bis heute. Weinheim: Juventa, S. 15-33.

Grimm, Kevin/Widaman, Keith (2010): Residual structures in latent growth curve modeling. In: Structural Equation Modeling 17, 3, S. 424-442. http://dx.doi.org/10.1080/10705511.2010.489006.

Kane, Michael (2011): The errors of our ways. In: Journal of Educational Measurement 48, 1, S. 12-30. http://dx.doi.org/10.1111/j.1745-3984.2010.00128.x.

Kelle, Udo (2008): Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. 2. Auflage. Wiesbaden: Springer VS. http://dx.doi.org/10.1007/978-3-531-91174-8.

Kline, Rex B. (2011): Principles and Practice of Structural Equation Modeling. 3. Auflage. New York: Guilford Press.

Kwok, Oi-Man/West, Stephen G./Green, Samuel B. (2007): The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models. A Monte Carlo Study. In: Multivariate Behavioral Research 42, 3, S. 557-592. http://dx.doi.org/10.1080/00273170701540537.

Lienert, Gustav A./Raatz, Ulrich (1998): Testaufbau und Testanalyse. 6. Auflage. Weinheim: Beltz.

Lord, Frederic M./Novick, Melvin R. (1968): Statistical Theories of Mental Test Scores with Contributions by A. Birnmaum. Reading: Addison-Wesley.

Marsh, Herbert W./Grayson, David (1994): Longitudinal stability of latent means and individual differences. A unified approach. In: Structural Equation Modeling 1, 4, S. 317-359. http://dx.doi.org/10.1080/10705519409539984.

Meredith, William/Tisak, John (1990): Latent curve analysis. In: Psychometrika 55, 1, S. 107-122. http://dx.doi.org/10.1007/BF02294746.

Newton, Paul E./Shaw, Stuart D. (2013): Standards for talking and thinking about validity. In: Psychological methods 18, 3, S. 301-319. http://dx.doi.org/10.1037/a0032969.

Nunnally, Jum C. (1978): Psychometric Theory. New York: McGraw-Hill.

Schurig, Michael (2017): Latente Variablenmodelle in der empirischen Bildungsforschung – Die Schärfe und Struktur der Schatten an der Wand. Dortmund: TU Dortmund. https://eldorado.tu-dortmund.de/bitstream/2003/36026/1/dissertation.pdf [Zugriff: 6. Februar 2018].

Schwippert, Knut (2016): Empirische Bildungsforschung: Perspektiven der Erziehungswissenschaft. In: Fickermann, D./Fuchs, H. W. (Hrsg.): Bildungsforschung – disziplinäre Zugänge. Fragestellungen, Methoden und Ergebnisse. Münster: Waxmann, S. 25-37.

Sivo, Stephen A./Fan, Xitao/Witta, Lea (2005): The biasing effects of unmodeled ARMA time series processes on latent growth curve model estimates. In: Structural Equation Modeling 12, 2, S. 215-231. https://doi.org/10.1207/s15328007sem1202_2.

Sivo, Stephen A./Willson, Victor L. (2000): Modeling causal error structures in longitudinal panel data. A Monte Carlo Study. In: Structural Equation Modeling 7, 2, S. 174-205. http://dx.doi.org/10.1207/S15328007SEM0702_3.

Skrondal, Anders/Rabe-Hesketh, Sophia (2004): Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models. Boca-Ranton: Chapman & Hall CRC. https://doi.org/10.1201/9780203489437.

Stevens, Stanley S. (1957): On the psychophysical law. In: Psychological Review 64, 3, S. 153-181. https://doi.org/10.1037/h0046162.

Steyer, Rolf/Eid, Michael (2001): Messen und Testen. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-56924-1.

Werner, Joachim (1997): Lineare Statistik. Das allgemeine lineare Modell. Weinheim: Psychologie Verl. Union Beltz.

Werner, Joachim (Hrsg.) (2005): Zeitreihenanalysen. Berlin: Logos.

Windzio, Michael/Teltemann, Janna (2013): Empirische Methoden zur Analyse kontextueller Faktoren in der Bildungsforschung. In: Becker, R./Schulze, A. (Hrsg.): Bildungskontexte. Wiesbaden: Springer Fachmedien, S. 31-60.

Zedler, Peter/Döbert, Hans (2010): Erziehungswissenschaftliche Bildungsforschung. In: Tippelt, R./Schmidt, B. (Hrsg.): Handbuch Bildungsforschung. 3. Auflage. Wiesbaden: Springer VS, S. 23-45.