Empirische Pädagogik
2015, 29. Jahrgang, Heft 4, S. 521-536

Nachhaltigkeit einer Lehrerfortbildung auf dem Prüfstand: Eine Replikationsstudie zur Gruppenrallye im Mathematikunterricht

Caroline Villiger, Alois Niggli, Christian Wandeler, Marcel Aebischer & Philippe Leopold

Das zentrale Interesse der vorliegenden Studie bestand in der Replikation von Ergebnissen einer Intervention zur Wirksamkeit der Gruppenrallye[1] im Mathematikunterricht von Fünftklässlern (Wandeler, Niggli, Villiger, Aebischer & Leopold, 2015). Untersucht wurde, ob Lehrkräfte, die in das besagte Forschungsprojekt involviert gewesen waren, nach zwei Jahren immer noch in der Lage waren, die Methode in ihren neuen Klassen mit vergleichbaren Lernerfolgen durchzuführen. Bei der ersten Studie war deutlich geworden, dass Mädchen signifikant schlechtere Leistungen zeigten. Deshalb wurde in der zweiten Studie zudem untersucht, ob sich dieser Effekt durch die Bildung geschlechtshomogener Gruppen abschwächen lässt. Alle Lehrkräfte waren in der Lage, die Gruppenrallye erneut durchzuführen. Im Posttest erreichten ihre Schüler wiederum höhere Lernergebnisse als die Kontrollgruppe. Die Signifikanz wurde jedoch knapp verfehlt. Ein positiver Effekt war im Follow-up-Test im Gegensatz zur ersten Intervention nicht mehr vorhanden. Diese Resultate werden in den Zusammenhang der Nachhaltigkeit von Fortbildungsmaßnahmen gestellt. Trotz der Bildung von geschlechtshomogenen Gruppen ließ sich der Einfluss des Geschlechts zudem nicht verringern. Die Förderung der Mädchen im Mathematikunterricht scheint umfassendere Maßnahmen zu verlangen.

Schlagwörter: Geschlechtshomogene Gruppen – Gruppenrallye – Mathematikleistung – Nachhaltigkeit von Lehrerfortbildung

1 Einleitung

Eine erste Studie zur Wirksamkeit der Gruppenrallyes im Mathematikunterricht von Fünftklässlern ist in Ausgabe 2/2015 dieser Zeitschrift bereits publiziert worden (Wandeler, Niggli, Villiger, Aebischer & Leopold, 2015). Es zeigte sich, dass selbst bei Kontrolle individueller Merkmale Kinder der Interventionsgruppe in einem curricularen Test signifikant besser abschnitten als Kinder einer vergleichbaren Kontrollgruppe. Dieser Effekt war unmittelbar nach der Intervention sowie bei einer Follow-up-Messung fünf Monate später erkennbar. Auf der Individualebene war vor allem der Einfluss des Geschlechts auffällig. Mädchen erreichten deutlich geringere Leistungen als Jungen. Beim vorliegenden Beitrag handelt es sich um eine Replikation dieser Studie. Dafür waren zwei Gründe ausschlaggebend. (1) In

[1] Bei der Gruppenrallye handelt es sich um eine Unterrichtsmethode, die einführenden Unterricht im Klassenverband, anschließende Gruppenarbeit, individuelle Leistungsüberprüfung und Gruppenbelohnung kombiniert.

erster Linie interessierte die Stabilität der Ergebnisse. Der Versuch wurde mit den beteiligten Lehrkräften zwei Jahre später wiederholt. (2) Bei dieser Gelegenheit wollten wir ferner versuchen, die Leistungen der Mädchen günstig zu beeinflussen.

1.1 Nachhaltigkeit von Fortbildung bei Lehrkräften

In der Forschung zur Fortbildung von Lehrkräften herrscht ein Mangel an Untersuchungen zur Nachhaltigkeit von Lernerfolgen. Effekte werden meist nur kurzzeitig, nach erfolgter Intervention erhoben (vgl. Timperley, 2008). Meist stehen lediglich Merkmale der Lehrkraft im Zentrum (Dennick, 2003; Watson, 2006) wie die Selbstwirksamkeit oder Skills zur Selbstregulation, ohne dass nach ihren Auswirkungen auf die Lernergebnisse der Schüler gefragt wird. Obwohl die Datenbasis über längerfristige Fortschritte daher als dünn gilt, scheinen zwei Aspekte besonders bedeutsam zu sein (Timperley, 2008): Die Nachhaltigkeit von kurzfristig erzielten Lernerfolgen hängt einerseits von den Erfahrungen ab, die während der professionsbezogenen Lernphase gemacht worden sind, andererseits aber auch von den organisatorischen Bedingungen, die gegeben sind, wenn der externe Support wegfällt. In der Lernphase haben sich nach Garet, Porter, Desimone, Birman und Yoon (2001) sowie Timperley, Wilson, Barrar und Fung (2007) folgende Kernfaktoren herauskristallisiert: In zeitlicher Hinsicht sollten Fortbildungen mindestens ein halbes Jahr dauern. Kontakte unter 14 Stunden innerhalb einer solchen Zeitspanne scheinen wirkungslos zu sein (Yoon, Duncan, Wen-Yu Lee, Scarloss & Shapley, 2007). Diese zeitlichen Voraussetzungen sind dann wirksam, wenn sie zu aktivem Lernen und gemeinsamer Partizipation beitragen. Der Fokus sollte ferner auf fachdidaktischem Wissen liegen und kohärent sein mit den curricularen Zielen im Unterricht. Neben oder nach den Fortbildungserfahrungen sollten in organisatorischer Hinsicht standortbezogene Stützmaßnahmen durch die Schulleitungen gegeben sein. Sie sollten die Wichtigkeit der Fortbildungsziele bekräftigen und die Lehrkräfte unterstützen, wenn sie Fortschritte überprüfen möchten oder externe Expertise wünschen (Franke, Carpenter, Fennema, Ansell & Behrend, 1998; Timperley, 2008).

1.2 Günstige Beeinflussung der Mathematikleistung von Mädchen durch geschlechtshomogene Gruppenbildung?

Im Kontrast zum wissenschaftlichen Konsens, wonach beide Geschlechter im Grundsatz über gleiche Begabungspotenziale in den mathematisch-naturwissenschaftlichen Fächern verfügen (Endepohls-Ulpe, 2012), konnten in Vergleichsstudien wiederholt Leistungsunterschiede zwischen Jungen und Mädchen festgestellt werden (Bos, Lankes, Prenzel, Schwippert & Walther 2004; Brehl, Wendt & Bos, 2011; Lehmann, Peek & Gänsfuss, 2011; im Längsschnitt: Wai, Cacchio, Putallaz & Makel, 2010). Auch wird im Zusammenhang mit mathematischen Leistungen oft

auf das vergleichsweise niedrige mathematische Selbstkonzept von Mädchen verwiesen (Faulstich-Wieland, 1991; Lehmann, 2006), das sich bereits im Primarschulalter abzeichnet (Eccles, Barber, Updegraff & O'Brien, 1998; Rustemeyer, 1998). In der vorangehenden Studie hatte das Geschlecht neben den Vorkenntnissen im Fach den zweithöchsten Einfluss auf die Leistungen der Schüler. Nach dem Verständnis des Design-based Research-Ansatzes (vgl. Reinmann, 2005) wurde deshalb versucht, die Bedingungen für die Mädchen zu optimieren, und zwar ohne das ursprüngliche kooperative Arrangement strukturell zu verändern. In Anlehnung an Argumente zur monoedukativen Bildung von Mädchen im Mathematikunterricht wurden geschlechtshomogene Rallyegruppen gebildet (Benölken, 2013). Zum einen erhoffte man sich, dass mit dieser Maßnahme bestehende Dominanzstrukturen zwischen den Geschlechtern aufgehoben und geschlechtsspezifische Zuschreibungen abgeschwächt würden (Graff, 2006). Dies schaffe größere Freiräume für die persönliche Entwicklung der Mädchen. Geschlechtshomogene Arbeitsgruppen könnten eine reibungslosere Zusammenarbeit ermöglichen und motivationsbeeinträchtigende Leistungsvergleiche zuungunsten der Mädchen hemmen. Weil vor allem mathematisch begabte Mädchen bei der Lösungserarbeitung zudem gern kooperativ arbeiten (Benölken, 2013), könnten insbesondere sie von homogenen Mädchengruppen profitieren. In den Metaanalysen von Ginsburg-Block, Rohrbeck und Fantuzzo (2006) sowie von Rohrbeck, Ginsburg-Block, Fantuzzo und Miller (2003) konnten denn auch höhere Effektstärken für Studien mit geschlechtshomogenen Gruppen für Lernleistung, Sozialkompetenz und Selbstkonzept registriert werden. Diese Befundlage spricht generell für eine geschlechtsspezifisch differenzielle Unterstützung und Förderung.

2 Fragestellungen

Ausgehend von den erwähnten Überlegungen liegt das Augenmerk in der vorliegenden Studie auf folgenden beiden Fragestellungen:

a) Können die positiven Effekte der Rallye-Methode mit denselben Lehrkräften auch zwei Jahre später (ohne erneute Fortbildung und Coaching) repliziert werden?

b) Kann die Entwicklung der Mathematikleistung von Mädchen durch geschlechtshomogene Gruppenbildung während der Intervention positiv beeinflusst werden?

3 Methode

3.1 Design

Die Nachfolgestudie wurde im Schuljahr 2012/2013 durchgeführt. Von den 15 Lehrkräften, die im Schuljahr 2010/2011 in Welle 1 die Workshops besucht hatten, konnten 11 Lehrkräfte für die zweite Studie nochmals rekrutiert werden. Vier Lehrkräfte hatten entweder die Stufe gewechselt oder ihre Stelle aufgegeben. Absagen waren infolgedessen keine zu verzeichnen. In einer Klasse wurde die Intervention von einer Praktikantin durchgeführt, der das Rallyekonzept aufgrund der Ausbildung an der Hochschule vertraut war. Sie wurde von der Rallye-Lehrkraft gecoacht. Eine weitere Lehrkraft reduzierte im Laufe des Jahres ihr Pensum. Die Rallyemethode wurde in dieser Klasse von der Nachfolgelehrkraft durchgeführt. Auch sie wurde von der ursprünglichen Lehrkraft gecoacht. In der Zeit zwischen den beiden Interventionen, im Schuljahr 2011/2012, hatte keine der beteiligten Lehrkräfte zusätzliche Erfahrungen mit der Gruppenrallye gesammelt. Ihre Klassen, die sie während zwei Jahren unterrichten, absolvierten 2012 ihr sechstes Grundschuljahr. Gegen Ende dieses Schuljahres ist von den Schülern eine Prüfung zum Übertritt in die Sekundarstufe I zu absolvieren. Lehrkräfte unterrichten in diesem Schuljahr eher stofforientiert und distanzieren sich von zuvor neu praktizierten methodischen Formen eher. Die Nachfolgestudie nahmen sie somit mit neuen Fünftklässlern in Angriff, die noch keine Erfahrungen mit der Gruppenrallye gemacht hatten. Dies entsprach ebenfalls den Bedingungen der ersten Welle. Diese Voraussetzung traf auch für die Modalitäten der Durchführung zu. Zu Beginn des Schuljahres setzten die Lehrkräfte wieder ein prototypisches Beispiel im Unterricht um, das von der Projektleitung ausgearbeitet worden war (s. Wandeler et al., 2015, S. 169). In einem einstündigen Briefing waren die organisatorischen Maßnahmen zuvor mit den Lehrkräften gemeinsam geklärt worden. Des Weiteren wurde die Bildung geschlechtshomogener Gruppen thematisiert. Eine fachliche Repetition zur Rallye erfolgte im Übrigen nicht mehr. Im März 2013 wurden die Daten der Vortests erhoben. Die Intervention mit den ursprünglichen drei Lernumgebungen wurde wiederum in den Monaten Mai und Juni durchgeführt. Die Posttests schlossen unmittelbar an die Intervention an. Ein Follow-up Test fand Mitte November 2013 statt. Die Tests wurden durch Testleiter administriert.

3.2 Stichprobe

Bei der Stichprobe handelte es sich um insgesamt 687 Fünftklässler aus dem deutschsprachigen Kantonsteil Freiburgs (Schweiz). Die Schüler stammten aus 41 Klassen und 22 Schulen. Die beteiligten Schüler setzten sich aus folgenden Gruppen zusammen: Interventionsgruppe 1: N = 262 (15 Klassen), Kontrollgruppe: N = 235 (15 Klassen) und Interventionsgruppe 2: N = 190 (11 Klassen). 48.5 % der

Schüler waren Mädchen, das Durchschnittsalter betrug 11.85 Jahre. 93.6 % der teilnehmenden Kinder waren in der Schweiz geboren und 77.9 % der Kinder gaben an, mit ihren Eltern deutsch zu sprechen. Die Zuteilung zu Intervention 1 und Kontrollgruppe geschah nach dem Zufallsprinzip (Rekrutierung der Lehrkräfte und ihre Klassen für die Intervention 2 siehe 3.1 Design). In Bezug auf ihre Zusammensetzung ergaben sich keine signifikanten Unterschiede zwischen den Gruppen.

3.3 Angaben zu den verwendeten Lernumgebungen

Für die Intervention wurden erneut die drei selben Lernumgebungen wie in Studie 1 (vgl. Wandeler et al., 2015) aus dem Lehrmittel „Schweizer Zahlenbuch 5" von Affolter, Armstad, Doebeli und Wieland (2009) ausgewählt. Diese Lernumgebungen bestehen aus Aufgaben, die hauptsächlich auf Texten und ikonischen Darstellungen basieren, und die in mathematische Modelle zu transformieren sind. Die Intervention erstreckte sich im Mittel über 15 Mathematik-Lektionen.

3.4 Treatmentcheck

Laut den Angaben der Lehrkräfte beanspruchte der Anteil selbstständigen kooperativen Lernens gemäß dem Rallyekonzept in der zweiten Interventionswelle durchschnittlich 55 % der Gesamtlernzeit. Er war damit praktisch identisch mit der dafür verwendeten kooperativen Lernzeit der ersten Welle (54.0 %). Der vorangehende Unterricht mit der Klasse betrug 42.3 % der Lernzeit (Welle 1: 43.0 %). Drei Prozent der Durchführungszeit wurden für andere Formen aufgewendet. Genannt wurden „Nachbesprechungen" und „Lernkontrollen verbessern". Innerhalb der Organisationsform „Klassenunterricht" wurden die Sozialformen wie in der ersten Welle ebenfalls variiert (43.3 % Plenumsunterricht, 22.5 % Einzelarbeit, 12.7 % Partnerarbeit, 21.3 % Gruppenarbeit). Die vorgesehenen drei Lernumgebungen konnten in allen Klassen vollumfänglich nach der Rallyemethode erarbeitet werden. In acht von 11 Klassen wurden Gruppen mit den positivsten Resultaten als Belohnung ein oder zwei Bonuspunkte für die individuelle Schlussprüfung gewährt (Welle 1: 10 von 15 Klassen). Gesamthaft gesehen ist in unterrichtsorganisatorischer Hinsicht zwischen den beiden Interventionen somit von sehr ähnlichen Durchführungsbedingungen auszugehen.

3.5 Messinstrumente

Mathematikleistungen: Zur Ermittlung relevanter Vorkenntnisse für die Bewältigung des Lernstoffs wurde ein standardisierter Vortest entwickelt. Ein weiterer curricularer Test wurde zum Lernstoff konstruiert, der in den drei Lernumgebungen zu behandeln war. Die Testskalierung wurde mit einem „3 parameter logistic item response theory model" mit der Software Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) durchgeführt. Der zum Messzeitpunkt T1 eingesetzte Test zur Erfas-

sung bedeutsamer Vorkenntnisse enthielt 19 Aufgaben mit 23 dichotomen Items. Der curriculare Test, der zu T2 und T3 (Follow-up) verwendet wurde, umfasste 18 Aufgaben mit 21 dichotomen Items. Die interne Konsistenz der beiden Tests (Kuder-Richardson-Formula 20) betrug .73 zu T1 und .74 zu T2 und zu T3.

SES: Der Sozioökonomische Status (SES) wurde mit dem ISEI-Index (International Socio-Economic Index of Occupational Status; Ganzeboom, De Graaf, Treiman & De Leeuw, 1992) bestimmt. Die Angaben der Kinder zum Beruf ihrer Mutter bzw. ihres Vaters wurden von zwei Ratern gesondert eingeschätzt. Nicht-Übereinstimmungen wurden kommunikativ validiert. Der höhere Score des jeweiligen Elternteils ging in die Analysen ein.

Familiensprache: Die Kinder wurden ebenfalls zur gesprochenen Sprache zu Hause befragt („Welches ist deine Muttersprache?"). Elf der am häufigsten gesprochenen Sprachen waren vorgegeben und konnten angekreuzt werden. Zudem konnte eine Kategorie „andere Sprachen" markiert werden. Die Variable wurde in eine Dummy-Variable (deutsch vs. andere Sprache) transformiert.

Intelligenz: Die kognitiven Fähigkeiten wurden mit dem sprachunabhängigen Intelligenztest CFT 20-R (Weiss, 1998) gemessen. Dieser Test enthält vier Subtests mit Aufgaben zu folgenden Bereichen: Reihenfortsetzen, Klassifikationen, Matrizen und topologische Schlussfolgerungen. Aus den vier Subtests wurde ein Gesamtscore gebildet (Cronbachs $\alpha = .59$).

Sprachliche Kompetenzen: Zur Erfassung des Leseverständnisses wurde der Lesetest ELFE 1-6 von Lenhard und Schneider (2006) verwendet. Eingesetzt wurde der Subtest, der das Leseverständnis auf Textebene misst (Auffinden von Informationen, satzübergreifendes Lesen, schlussfolgerndes Denken). Die interne Konsistenz (Kuder-Richardson-Formula 20) betrug $\alpha_{T1} = .80$ und $\alpha_{T2} = .79$.

Zusammenarbeit in der Gruppe: Zur Erfassung der Zusammenarbeit in der Gruppe wurden die Interventionsgruppen 1 und 2 während der Intervention dreimal auf einer vierstufigen Skala befragt. Die Items wurden zum Teil Huber (2007) entnommen (z. B. „Ich habe mich bei der Zusammenarbeit wohl gefühlt") und durch weitere Items ergänzt (z. B. „In der Gruppe haben wir uns gegenseitig geholfen"). Die Skala wies eine Reliabilität von Cronbachs $\alpha_{t1, t2} = 0.83$ und $\alpha_{t3} = .88$ auf.

3.6 Statistisches Vorgehen

Aufgrund der hierarchischen Datenstruktur wurden zur Überprüfung der Hypothesen Mehrebenenanalysen durchgeführt (Raudenbush & Bryck, 2002). Dazu wurde das Softwarepaket HLM 6.04 verwendet (Raudenbush, Bryck, Cheong & Congdon, 2004), welches allein unstandardisierte Koeffizienten berichtet. Zur vereinfachten Interpretation von Effekten der Prädiktorvariablen wurden alle met-

rischen Variablen z-standardisiert (M = 0, SD = 1), was bei der Berechnung einer y-Standardisierung gleichkommt. Die Regressionskoeffizienten zeigen dann an, um welchen Anteil einer Standardabweichung sich die abhängigen Variablen bei der Zu- oder Abnahme einer Prädiktorvariable um eine Standardabweichung verändern. Die deskriptiven Analysen wurden mit dem Statistikpaket SPSS 17 gerechnet.

Der Durchschnitt der fehlenden Werte lag bei 1.88 % pro Variable. Aufgrund der niedrigen Anzahl an Ausfällen und der Annahme, dass die Daten völlig zufällig fehlten (engl. missing completely at random, MCAR; vgl. Little & Rubin, 2002), wurden die MDM-files (HLM 6.04, Raudenbush et al., 2004) aufgrund der Rohdaten erstellt. Bei den Analysen wurde folglich bei fehlenden Daten der fallweise Ausschluss angewendet (engl. listwise deletion). Für die Berechnung der Effektstärke wurde die Formel für dichotome Variablen in Mehrebenenmodellen verwendet (vgl. Tymms, Merrell & Henderson, 1997; Wandeler et al., 2015).

4 Ergebnisse

4.1 Deskriptive Analysen und Interkorrelationen der Variablen

In Tabelle 1 werden Mittelwerte, Standardabweichungen und Interkorrelationen der Variablen berichtet. Zu beachten ist, dass sich der Mathematiktest stoffrelevanter Vorkenntnisse zu T1 inhaltlich vom curricularen Test zu T2 und T3 unterscheidet, mit dem die in den drei Lernumgebungen erreichten Leistungen gemessen worden sind. Das Geschlecht (Jungen = 1) korrelierte negativ mit dem Textverständnis, jedoch positiv mit den Mathematikleistungen. Die Korrelationen der Leistungstests mit dem beruflichen Status der Eltern waren höher als diejenigen mit der Familiensprache Deutsch. Die Korrelationen zwischen den Mathematiktests und dem sprachunabhängigen Intelligenztest fielen stärker aus als die Zusammenhänge mit dem Textverständnis. Auffallend hoch waren die Interkorrelationen der drei Mathematiktests.

Tabelle 1: Mittelwerte, Standardabweichungen und Korrelationen

	M	SD	1	2	3	4	5	6	7	8	9	10
1 Geschl. (1 = M)	--[a]	--[a]										
2 Deutsch (= 1)	--[a]	--[a]	.05									
3 ISEI Eltern	50.05	17.00	-.05	.12								
4 CFT	36.11	5.93	-.07	.03	.17							
5 ELFE-Textverst.	15.93	3.21	-.14	.12	.19	.21						
6 Mathetest T1	10.28	4.17	.11	.17	.22	.44	.40					
7 Mathetest T2	9.72	3.99	.18	.18	.22	.46	.40	.69				
8 Mathetest T3	10.40	3.92	.18	.15	.19	.42	.38	.68	.74			
9 Interv. 1 (= 1)	--[a]	--[a]	.01	.05	-.01	-.05	.05	.03	.07	.08		
10 Interv. 2 (= 1)	--[a]	--[a]	-.01	-.14	.01	.07	-.09	-.09	-.03	-.07	-.47	
11 KG	--[a]	--[a]	-.01	.08	.00	-.01	.04	.05	-.04	-.02	-.58	-.45

Anmerkungen: [a] dichotome Variablen; r > .08 = p < .05; r > .11 = p < .01; r > .13 = p < .001

In Tabelle 2 sind die Mittelwerte und Standardabweichungen der Mathematiktests für die drei Probandengruppen zu den drei Messzeitpunkten aufgeführt. Aufgrund der unterschiedlichen Tests zu T1 und T2/T3 können keine interpretativen Aussagen gemacht werden zum Leistungszuwachs zwischen T1 und T2. Der Mittelwert der Interventionsgruppe 2 (IG 2; geschlechtshomogene Gruppen) war im Vortest tiefer, jedoch besteht gemäß Scheffé Test kein signifikanter Unterschied zu den beiden anderen Gruppen (IG 1 und Kontrollgruppe). Die Mittelwerte der Interventionsgruppe 2 waren beim Nachtest und Follow-up beide Male tiefer als bei der Interventionsgruppe 1. Zudem war der Leistungszuwachs zwischen diesen beiden Messzeitpunkten geringer. In Bezug auf die geschlechtsspezifischen Daten zeigte sich, dass die Geschlechterunterschiede in der Interventionsgruppe 2 im Vortest größer waren als in der Interventionsgruppe 1. Dieser Befund war ebenfalls für den Nachtest zu erkennen (Unterschied Mädchen-Jungen bei der IG 2: knapp eine halbe Standardabweichung). Beim Follow-up war der Geschlechterunterschied bei der IG 2 hingegen geringer.

Tabelle 2: Mittelwerte und Standardabweichungen der Probandengruppen im Vortest und in den curricularen Nachtests

	IG 1		IG 2		KG	
	M	SD	M	SD	M	SD
Vortest (gesamt)	**10.53**	**4.08**	**9.72**	**4.18**	**10.60**	**4.29**
Jungen	10.81	4.02	10.34	4.30	11.11	4.38
Mädchen	10.21	4.04	9.08	3.99	10.19	4.14
Nachtest (gesamt)	**10.13**	**3.91**	**9.51**	**4.05**	**9.51**	**4.06**
Jungen	10.67	3.91	10.46	4.05	10.15	4.14
Mädchen	9.49	3.77	8.54	3.94	8.80	3.86
Follow up (ges.)	**10.89**	**3.82**	**9.99**	**4.02**	**10.32**	**3.96**
Jungen	11.55	3.83	10.50	3.95	11.13	4.00
Mädchen	10.06	3.65	9.44	4.03	9.47	3.72

Des Weiteren wurde für die Replikationsstudie die Zusammenarbeit in der Gruppe untersucht. Vor dem Hintergrund der zweiten Fragestellung interessierte, ob die Mädchen der Interventionsgruppe 2 ihre Gruppe bzw. die Arbeitsatmosphäre in der Gruppe positiver einschätzten als die Mädchen der Interventionsgruppe 1. Entgegen der Erwartungen konnten jedoch keine signifikanten Unterschiede festgestellt werden ($t(220) = 1.85$; $p = .06$). Die Mädchen der Interventionsgruppe 2 schätzten die Gruppenzusammenarbeit im Gegenteil sogar etwas weniger positiv ein als die Mädchen der Gruppe 1.

4.2 Vorhersage der Leistungsentwicklung

In Tabelle 3 sind die vier Modelle zur Vorhersage der Mathematikleistung zum Zeitpunkt T2 dargestellt.

In Modell 1 wurden lediglich die mathematischen Vorkenntnisse als Prädiktor verwendet. In Modell 2 wurde auf der Klassenebene der Einfluss der beiden Interventionen 1 und 2 untersucht (Referenz: Kontrollgruppe). Das Modell 3 prüft zusätzlich den Einfluss weiterer relevanter Kovariaten auf der Individualebene: das Geschlecht des Kindes, die Erstsprache, der berufliche Status der Eltern (ISEI), sowie die nonverbale Intelligenz (CFT) und das Textverständnis des Kindes (ELFE).

Der Effekt der Intervention 1 auf die Mathematikleistung zum Zeitpunkt T2 war statistisch signifikant und blieb über die Modelle hinweg konstant (B = .16,

p < .05). Der Effekt der Intervention 2 war positiv, jedoch nicht statistisch signifikant (p = .07) und sank unter Kontrolle der Kovariaten. Eine Veränderung der Referenzgruppe (direkter Vergleich der Interventionen 1 und 2) im Rahmen von weiteren Analysen ergab zudem keinen signifikanten Unterschied zwischen den beiden Interventionsgruppen. Die Individualprädiktoren (Geschlecht, Erstsprache, ISEI, nonverbale Intelligenz und Textverständnis) waren im Modell M3 alle statistisch signifikant und erhöhten die aufgeklärte Varianz um .06.

In Modell 3 betrugt die Effektstärke für die Intervention 1 Δ = .24 und für die Intervention 2 Δ = .17 (vgl. Tymms, Merrell & Henderson, 1997). Die Effektstärke war für beide Interventionen bedeutsam, jedoch nur für die Intervention 1 statistisch signifikant.

Tabelle 3: Befunde (Regressionskoeffizienten) aus Mehrebenenanalysen zur Vorhersage der Mathematikleistung zum Zeitpunkt T2

		M1			M2			M3		
		B	SE	p	B	SE	p	B	SE	p
Individual-ebene	Vorkenntnisse T1	.70	.02	<.000	.70	.02	.000	.49	.03	<.000
	Geschlecht (m)							.34	.04	<.000
	Erstspr. Deutsch							.14	.06	.021
	ISEI							.07	.03	.011
	CFT							.21	.03	<.000
	Textverst. ELFE							.15	.03	<.000
Klassen-ebene	IG 1				.16	.08	.030[a]	.16	.07	.015[a]
	IG 2				.14	.09	.066[a]	.11	.09	.101[a]
	KG									
	R^2	.48			.49			.55		

Anmerkungen: [a] der gerichteten Hypothese entsprechend, p-Werte für einseitigen Signifikanztest

In Tabelle 4 werden die Ergebnisse zur Vorhersage der Mathematikleistung zum Messzeitpunkt T3 berichtet. Während für die Intervention 1 selbst zu diesem Zeitpunkt noch Effekte nachweisbar waren, und dies konstant über die Modelle, stellte sich der Effekt der Intervention 2 als deutlich abgeschwächt heraus (M2: B = .05 und M3: B = .03). Weiter ist zu beachten, dass die Variablen des familiären Hintergrundes (Erstsprache, Berufsstatus) zu T3 nicht mehr statistisch signifikant waren.

Zum Messzeitpunkt T3 lag die Effektstärke der Intervention 1 bei Δ = .25 und für die Intervention 2 bei Δ= .05. Für beide Messzeitpunkte konnten zudem keine Interaktionseffekte zwischen den Vorkenntnissen der Schüler und der Intervention festgestellt werden. Schwächere und stärkere Schüler profitierten somit von der

Intervention in ähnlichem Ausmaße. Derselbe Befund konnte für das Geschlecht der Schüler registriert werden.

Tabelle 4: Befunde (Regressionskoeffizienten) aus Mehrebenenanalysen zur Vorhersage der Mathematikleistung zum Zeitpunkt T3

		M1			M2			M3		
		B	SE	p	B	SE	p	B	SE	p
Individual-ebene	Vorkenntnisse T1	.68	.02	<.000	.68	.02	.000	.50	.03	<.000
	Geschlecht (m)							.37	.06	<.000
	Erstspr. Deutsch							.09	.06	.145
	ISEI							.04	.03	.148
	CFT							.18	.03	<.000
	Textverst. ELFE							.15	.03	<.000
Klassen-ebene	IG 1				.16	.09	.033 [a]	.17	.08	.019 [a]
	IG 2				.05	.09	.263 [a]	.03	.09	.363 [a]
	KG									
	R^2	.45			.46			.52		

Anmerkungen: [a] der gerichteten Hypothese entsprechend, p-Werte für einseitigen Signifikanztest

5 Diskussion

Mit der vorliegenden Replikationsstudie zu Wandeler et al. (2015) wurde beabsichtigt, Effekte einer Intervention zwei Jahre später nochmals zu prüfen. Die Wirksamkeit der Gruppenrallye konnte mit den ursprünglich involvierten Lehrkräften an einer neuen Schülerstichprobe im Posttest nicht repliziert werden. Immerhin war aber ein positiver Effekt der Intervention 2 nachweisbar, wenn auch nicht statistisch signifikant. Aufschlussreich ist der Befund, dass in einem direkten Vergleich beider Interventionsgruppen beim Posttest keine statistisch signifikanten Unterschiede nachgewiesen werden konnten. Dies bedeutet, dass die Intervention beim zweiten Durchgang ähnlich erfolgreich war. Der positive Effekt war zum Zeitpunkt der Follow-up Messung, fünf Monate später, im Gegensatz zur ersten Intervention hingegen vollständig verschwunden. Die zweite Intervention zeigte demnach offensichtlich keine nachhaltigen Effekte. In der Folge werden diese Befunde im Detail diskutiert.

Vorab ist anzumerken, dass es den Lehrkräften gelungen ist, die Methode im Unterricht erneut umzusetzen (siehe dazu auch die Angaben zum Treatmentcheck). Die Verbindung von gemeinsamem Unterricht mit der Klasse zur Erarbeitung der Grundlagen, selbständigem Weiterlernen in den Gruppen, regelmäßigem Feedback

und Belohnungsmaßnahmen scheint kurzfristig erneut wirksam gewesen zu sein. Bei diesen Elementen handelt es sich um organisatorische Maßnahmen der Gruppenrallye, die eher die Makroebene des Unterrichts strukturieren. Nachhaltigere Wirkungen dürften jedoch von den Interaktionsprozessen ausgehen, die sich auf der Mikroebene des Lernens in den Gruppen abspielen (Lipowsky, 2006). Diese werden in einer Rallye durch eine Anleitung (Script) moderiert, die den Gruppen zur Verfügung steht. Sie soll beim Lernen gegenseitige Wechselwirkungen und Feedbacks anregen, die dazu beitragen, kognitive Verknüpfungen von neuem mit bestehendem Wissen herzustellen, das Verständnis zu überwachen und bei Schwierigkeiten einander zu helfen (Renkl, 1997). Auch in Gruppen kann von einzelnen jedoch mit mehr oder weniger Interesse verfolgt werden, was gemeinsam bearbeitet werden soll. In solchen Fällen ist die passende Einwirkung der Lehrkraft unerlässlich. In der ersten Intervention waren die Lehrkräfte im Rahmen der Fortbildungsworkshops herausgefordert gewesen, sich mit solchen Prozessen zu befassen und sich über die gemachten Erfahrungen in ihrer Klasse auszutauschen. Möglicherweise achteten sie in der zweiten Intervention weniger auf das Interaktionsgeschehen in den Gruppen und konzentrierten sich auf die konstituierenden Elemente der Methode, die von außen plan- und steuerbar sind. Vermutlich waren den Lehrkräften zwei Jahre später vor allem die praktischen Prozeduren auf der Oberfläche (oder Sichtstruktur) des Lernens präsent. Sie könnten daher die kooperativen Prinzipien, die hinter den eingeleiteten Maßnahmen stehen, weniger intensiv beachtet haben. Erst die Orientierung an solchen Prinzipien und nicht lediglich die Umsetzung oberflächlicher Prozeduren lässt jedoch nachhaltige Wirkungen wahrscheinlicher werden (Brown & Campione, 1996).

Keinen Einfluss auf die Leistungsentwicklung der Mädchen hatte der Versuch, geschlechtshomogene Rallyegruppen zu bilden. Dieses Ergebnis ist einerseits durch den Umstand erklärbar, dass lediglich ein organisatorisches Oberflächen-Element umgesetzt worden war: Die Bildung geschlechtshomogener Gruppen. Die spezifische Förderung der Mädchen im Fach Mathematik dürfte dagegen umfassendere Maßnahmen verlangen wie beispielsweise angepasstes Coaching oder auch fachliche Differenzierungselemente (Benölken, 2013). Zudem war die Gruppenzusammenarbeit von den Mädchen der Intervention 2 sogar etwas weniger positiv eingeschätzt worden als von den Mädchen der Intervention mit geschlechtsheterogenen Gruppen. Deshalb ist rückblickend davon auszugehen, dass letztere von den Jungen nicht in einem Ausmaß dominiert worden sind, welches ihr Lernen entscheidend beeinträchtigt hätte. Zumindest hatte sich die Lernsituation in geschlechtshomogenen Gruppen im Vergleich zu heterogenen Gruppen aus der Sicht der Mädchen eher etwas verschlechtert.

Limitierende Aspekte der Untersuchung sind bei Wandeler et al. (2015) im Detail aufgeführt. Für die Replikationsstudie ist bedeutsam, dass sich vier Lehrkräfte nicht mehr beteiligen konnten. Weil es sich dabei um natürliche Fluktuationen in der Lehrerschaft handelt und keine Absagen entgegenzunehmen waren, sind systematische Effekte wenig wahrscheinlich. Im Weiteren könnte eine Konfundierung der Effekte bemängelt werden, weil in der zweiten Intervention geschlechtshomogene Rallye-Gruppen gebildet worden sind. Streng genommen wären die beiden Treatments somit nicht vergleichbar. Diese mögliche Konfundierung wurde durch die Prüfung der Interaktion von Geschlecht und Methode kontrolliert und hat sich als unerheblich herausgestellt. Ferner ist es für Design-based Ansätze charakteristisch, implementierte Praxis mit neuen gerichteten Hypothesen zu optimieren (Brown, 1992).

Aus der Sicht der Praxis sollten sich Veränderungen nicht nur kurzfristig, sondern auch auf Dauer als fruchtbar herausstellen. Unsere Studie liefert dazu einige Impulse. Fortbildungen von ausreichender Dauer, die aktives Lernen beinhalten und die auf die curricularen Bedingungen in der Praxis abgestimmt sind, können offensichtlich von den Lehrkräften auch nach zwei Jahren erfolgreich umgesetzt werden. Allerdings beschränkte sich der positive Effekt auf den Kenntnisstand der Schüler unmittelbar nach dem Treatment. Nachhaltigere Wirkungen sind vermutlich dann zu erwarten, wenn neben den Prozeduren auf der Oberfläche auch Prozessen auf der Mikroebene des Lernens ausreichend Beachtung geschenkt wird (Kunter & Trautwein, 2013). Diese verlangen nach Reflexion und Austausch zwischen den beteiligten Lehrkräften. Zu diesem Zweck dürften standortbezogene Stützmaßnahmen oder generell Auffrischungssitzungen unumgänglich sein. Im Hinblick auf die spezifische Förderung der Mädchen im Fach Mathematik scheinen ferner isolierte Maßnahmen, die in der Praxis umgesetzt werden, wenig erfolgreich zu sein. Erfolgversprechender dürfte eine fachdidaktische Gesamtstrategie sein. Als Fazit der Replikationsstudie lässt sich somit festhalten: Nur wenn eine reflektierte Praxis implementiert werden kann, die Prozessen auf der Mikroebene des Lernens Beachtung schenkt, scheinen Fortbildungsanstrengungen auch längerfristig lernwirksam für die Schüler zu sein. Dazu wäre es notwendig, für die beteiligten Lehrkräfte Begleitmaßnahmen in Form von Austauschangeboten oder zumindest Selbstkontrollmöglichkeiten bereitzustellen.

Literatur

Affolter, W., Amstad, H., Doebeli, M. & Wieland, G. (2009). Schweizer Zahlenbuch 5. Zug: Klett und Balmer Verlag.

Benölken, R. (2013). Geschlechtsspezifische Besonderheiten in der Entwicklung mathematischer Begabungen. Mathematica didactica, 36, 66-96.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K. & Walther, G. (2004). IGLU: Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. Münster: Waxmann.

Brehl, T., Wendt, H. & Bos, W. (2011). Geschlechtsspezifische Unterschiede in mathematischen und naturwissenschaftlichen Kompetenzen. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.). Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich (S. 203-230). Münster: Waxmann.

Brown, A. L. (1992). Design Experiments: Theoretical and methodological challenges in creating complex intervention in classroom settings. The Journal of the Learning Sciences, 2 (2), 141-178.

Brown, A. & Campione, J. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.). Innovations in learning (p. 289-326). Hillsdale, NJ: Lawrence Erlbaum.

Dennick, W. (2003). Long-term retention of teaching skills after attending the Teaching Improvement Project: a longitudinal, self-evaluation study. Medical Teacher, 25 (3), 314-318.

Eccles, J., Barber, B. L., Updegraff, K. & O'Brien, K. M. (1998). An expectancy-value model of achievement choices: The role of ability self-concepts, perceived task utility and interest in predicting activity choice and course enrollment. In L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Eds.). Interest and learning (p. 267-279). Kiel: IPN.

Endepohls-Ulpe, M. (2012). Begabte Mädchen und Frauen. In H. Stöger, A. Ziegler & M. Heilemann (Hrsg.). Mädchen und Frauen in MINT. Bedingungen von Geschlechtsunterschieden und Interventionsmöglichkeiten (S. 103-132). Berlin: LiT-Verlag.

Faulstich-Wieland, H. (1991). Koedukation - enttäuschte Hoffnungen? Darmstadt: Wissenschaftliche Buchgesellschaft.

Franke, M. L., Carpenter, T., Fennema, E., Ansell, E. & Behrend, J. (1998). Understanding teachers' self-sustaining, generative change in the context of professional development. Teaching and Teacher Education, 14 (1), 67-80.

Ganzeboom, H. B. G., De Graaf, P. M., Treiman, D. J. & De Leeuw, J. (1992). A standard international socio-economic index of occupational status. Social Science Research 21 (1), 1-56.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F. & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38(4), 915-945.

Ginsburg-Block, M. D., Rohrbeck, C. A. & Fantuzzo, J. W. (2006). A meta analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. Journal of Educational Psychology, 98 (4), 732-749.

Graff, U. (2006). Tough enough to wear pink! Impulse der neuen Geschlechterdebatte in der Pädagogik. In K. Böllert (Hrsg.). Von der Delegation zur Kooperation. Bildung in Familie, Schule, Kinder- und Jugendhilfe (S. 85-94). Wiesbaden: VS Verlag für Sozialwissenschaften.

Huber, A. A. (2007). Wechselseitiges Lehren und Lernen (WELL) als spezielle Form kooperativen Lernens. Berlin: Logos Verlag.

Kunter, M. & Trautwein, U. (2013). Psychologie des Unterrichts. Paderborn: Schöningh.

Lehmann, R. H. (2006). Mädchen und Mathematik in der gymnasialen Sekundarstufe I - Ergebnisse einer Längsschnittstudie. In I. Hosenfeld & F.-W. Schrader (Hrsg.). Schulische Leistung - Grundlagen, Bedingungen, Perspektiven (S. 107-120). Münster: Waxmann.

Lehmann, R. H., Peek, R. & Gänsfuss, R. (2011). LAU 5 Aspekte der Lernausgangslage und der Lernentwicklung - Klassenstufe 5. In Behörde für Schule und Berufsbildung (Hrsg.). LAU - Aspekte der Lernausgangslage und der Lernentwicklung (S. 15-119). Münster: Waxmann.

Lenhard, W. & Schneider, W. (2006). ELFE 1-6. Ein Leseverständnistest für Erst- bis Sechstklässler. Göttingen: Hogrefe.

Lipowsky, F. (2006). Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann-Ghionda & E. Terhart (Hrsg.). Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern (S. 47-70). Weinheim: Beltz.

Little, R. J. A. & Rubin, D. B. (2002). Statistical analysis with missing data. New York: Wiley.

Raudenbush, S. W. & Bryck, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryck, A. S. Cheong, Y. F. & Congdon, R. (2004). HLM 6: Hierarchical linear and nonlinear modeling. Chicago, IL: Scientific Software International.

Reinmann, G. (2005). Innovation ohne Forschung. Ein Plädoyer für den Design-Based Research-Ansatz in der Lehr-Lernforschung. Unterrichtswissenschaft, 33(1), 52-69.

Renkl, A. (1997). Lernen durch Lehren. Zentrale Wirkmechanismen beim kooperativen Lernen. Wiesbaden: Deutscher Universitäts-Verlag.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W. & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. Journal of Educational Psychology, 95(2), 240-256.

Rustemeyer, R. (1998). Geschlechtsstereotype und ihre Auswirkungen auf das Sozial- und Leistungsverhalten. Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 8, 115-129.

Timperley, H. (2008). Teacher professional learning and development. Geneva: International Academy of Education and International Bureau of Education (IBE).

Timperley, H., Wilson, A., Barrar, H. & Fung, I. (2007). Teacher professional learning and development: Best evidence synthesis iteration (BES). Ministery of Education. Wellington: New Zealand.

Tymms, P., Merrell, C. & Henderson, B. (1997). The first year at school: A quantitative investigation of the attainment and progress of pupils." Educational Research and Evaluation, 3(2), 101-118.

Wai, J., Cacchio, M., Putallaz, M. & Makel, C. (2010). Sex differences in the right tail of cognitive abilities: A 30 years examination. Intelligence, 38(4), 412-423.

Wandeler, C., Niggli, A., Villiger, C., Aebischer, M. & Leopold, P. (2015). Ein Quasi-Experiment zur Gruppenrallye im Mathematikunterricht: Hält die Methode, was sie verspricht? Empirische Pädagogik, 29(2), 161-188.

Watson, G. (2006). Technology Professional Development: Long-term effects on teacher self-efficacy. Journal of Technology and Teacher Education, 14(1), 151-165.

Weiss, R. H. (1998). Grundintelligenztest Skala 2: CFT 20. Braunschweig: Westermann.

Yoon, K. S., Duncan, T., Wen-Yu Lee, S., Scarloss, B. & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. U. S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software] Skokie, IL: Scientific Software International, Inc.

Sustainability of a teacher training put to the test: A replication study on STAD in mathematics

The main interest of the present study was to replicate the results of an intervention study on the effectiveness of STAD (student teams achievement devision) in mathematics teaching of 5th-graders (Autoren, 2015). We investigated whether teachers, who had been involved in the said research project, after two years were still able to carry out the method in their new classes and obtain comparable learning outcomes. In the first study, it was found that girls showed significantly poorer performance. Therefore, the second study examines whether this effect can be reduced through the formation of gender-homogeneous grouping. All teachers have been able to carry out the STAD again. In the posttest their students again achieved higher learning outcomes than the control group, but the difference was not statistically significant. Unlike in the first intervention, a positive effect was no longer present in the follow-up test. These results are discussed in the context of the sustainability of teacher training activities. Despite the same-gender grouping, the influence of gender did not decrease. The encouragement of girls in mathematics education seems to require more extensive measures.

Keywords: mathematics performance – same-gender grouping – STAD student-teams-achievement devision – sustainability of teacher training

Autoren

Dr. Caroline Villiger,
Prof. Dr. Alois Niggli, Pädagogische Hochschule Freiburg, Freiburg/CH,
Prof. Dr. Christian Wandeler, California State University, Department of Curriculum and Instruction, Fresno/CA,
Marcel Aebischer, Praxisdozent,
Philippe Leopold, M.A., Pädagogische Hochschule Freiburg, Freiburg/CH.
Korrespondenz an: villigerc@edufr.ch