Erschienen in: Lehren und Lernen ; 40 (2014), 2. - S. 15-19

Birgit Mirjam Beisswingert, Thomas Götz

Power of Difference – Wissenschaftliche Begleitung und Evaluation

Die Bildungslandschaft ist geprägt von stetigen Veränderungs- und Optimierungsbemühungen. Wie Schülerinnen und Schüler in unseren Schulen bestmöglich gefördert werden können und welche – auch strukturellen – Veränderungen zur besseren Nutzung und Förderung der individuellen Potenziale unserer Kinder und damit zu optimalen Lernergebnissen beitragen können, ist ein Thema ständigen schulpolitischen, aber auch gesellschaftlichen Diskurses.
Aktuelle prominente Beispiele sind hier etwa Überlegungen zum angemessenen Umgang mit Heterogenität unter den Schülern, struktur- und prozessbezogene Entwicklungen unter dem Stichwort „Inklusion", die Einführung von Gemeinschaftsschulen sowie anhaltende Diskussionen über das acht- bzw. neunjährige Gymnasium. Daneben gibt es auch viele innovative Ideen, die in klein angelegten Modellprojekten auf ihre Umsetzbarkeit hin getestet werden, wie etwa das Projekt „Power of Difference" am Gymnasium Wilhelmsdorf.

Wissenschaftliche Begleitung und Evaluation – Wozu?

Stehen dann jedoch Entscheidungen über die Fortsetzung des Modellprojekts an, über eine Ausweitung auf weitere Schulen, oder sollen möglicherweise sogar gesetzliche Änderungen erwirkt werden, so ist dies häufig mit Fragen nach den Effekten der innovativen Interventionen verbunden. Derartige evidenzbasierte gesellschaftliche und bildungspolitische Diskurse gehen über Auseinandersetzungen vor dem Hintergrund individueller Überzeugungen („Ideologien") und pädagogisch-psychologischer Theorietraditionen hinaus, indem sie in der Lage sind, mittels empirischer Befunde Empfehlungen ableiten und dadurch zu wissenschaftlich fundierten schulpolitischen Entscheidungen beitragen zu können.

Wenn verlässliche Aussagen über die Wirksamkeit von Interventionen, also Veränderungen im Schulalltag durch die Durchführung von Modellprojekten, getroffen werden sollen, dann sind exakte wissenschaftliche Herangehensweisen notwendig. Nur dadurch kann sichergestellt werden, dass die Ergebnisse der Wirksamkeitsuntersuchungen (Evaluation) tatsächlich auf die Veränderungen im schulischen Alltag zurückgeführt und zulässige Schlüsse aus den Ergebnissen gezogen werden können. Das gesamte Projekt – von seiner Planungs- über die Durchführungs- und Auswertungsphase – muss daher in bestimmter Weise (z. B. unter Kontrolle möglichst vieler Einflussfaktoren, die die wissenschaftliche Evaluation potenziell stören könnten) durchgeführt werden, damit abschließende wissenschaftlich fundierte Aussagen über mögliche Effekte der Interventionen überhaupt zulässig sind.

In diesem Sinne wurde das Projekt „Power of Difference" von Beginn seiner Durchführung an von Mitarbeiter/-innen der Arbeitsgruppe „Empirische Bildungsforschung" der Universität Konstanz/Pädagogischen Hochschule Thurgau begleitet. Die Ziele der wissenschaftlichen Begleitung lagen darin, das verantwortliche Koordinationsteam vor Ort am Gymnasium Wilhelmsdorf dabei zu unterstützen, das Projekt so durchzuführen, dass möglichst optimale Rahmenbedingungen für eine fundierte wissenschaftliche Evaluation der Projekteffekte gegeben waren. Außerdem wurde – in Absprache mit den Verantwortlichen an der Schule – die wissenschaftliche Evaluation von Grund auf geplant und durchgeführt, um solide Aussagen über die Auswirkungen der Intervention treffen zu können. Diese wissenschaftliche Begleitung umfasst dabei verschiedene Schritte, die in den folgenden Abschnitten jeweils in allgemeiner Form sowie in ihrer Anwendung auf das Projekt „Power of Difference" dargestellt werden.

Was ist eine wissenschaftliche Evaluation?

Im Zentrum einer wissenschaftlichen Evaluation steht häufig die Frage nach der Wirksamkeit einer Intervention; es handelt sich dabei also oft um Wirksamkeitsforschung. Die Intervention soll bei den an der Intervention Beteiligten (Interventionsgruppe) eine Veränderung in Richtung eines definierten Ziels bewirken, während diese Veränderung in einer alternativen (bzw. nicht vorhandenen) Maßnahme (Kontrollgruppe) nicht beobachtbar sein sollte (Hager 2008; Wild & Möller 2009). Auf der Basis dieser theoretischen Überlegungen mussten also für die Planung der Evaluationsstudie folgende Aspekte definiert werden:

- **Intervention**
  Welche Eingriffe in den schulischen Alltag werden – im Vergleich zum vorherigen Normalzustand – durch das Projekt „Power of Difference“ vorgenommen? Konkret mussten beispielsweise die spezifischen Aufgabengebiete und Einsatzfelder der Unterrichtsassistenten sowie der Förderpädagogen/Coaches definiert werden.
- **Veränderung**
  Um Aussagen über Veränderungen treffen zu können, müssen die Ist-Zustände zu mindestens zwei verschiedenen Zeitpunkten, nämlich vor Beginn sowie nach Abschluss der Intervention erfasst, also sogenannte Vorher-Nachher-Messungen geplant werden.
- **Interventions- versus Kontrollgruppe**
  Die Zielgruppe der Interventionen musste definiert werden, um für die Datenerhebungen sowohl eine von der Intervention betroffene Stichprobe (Interventionsgruppe) als auch eine davon nicht betroffene Stichprobe (Kontrollgruppe) festlegen zu können, deren Entwicklungen zwischen der Vorher-Nachher-Messung miteinander verglichen werden können.
- **Definiertes Ziel der Intervention**
  Anhand der bereits definierten Intervention, also der Aufgabengebiete und Einsatzfelder des in „Power of Difference“ zusätzlichen Personals, wurde vor dem Hintergrund wissenschaftlicher Befunde ein theoretisches Rahmenmodell entwickelt, das Annahmen darüber zuließ, wie sich die Interventionen auf unterschiedliche Variablen und die verschiedenen am Projekt beteiligten Akteure im Schulleben auswirken sollte. Daraus konnten die Zielsetzungen der Interventionen sowie die in der Evaluation zu untersuchenden interessierenden Variablen abgeleitet werden.

Methodik der wissenschaftlichen Evaluation: Evaluationsdesign

Auf Basis der nun definierten Aspekte konnte ein konkretes Evaluationsdesign festgelegt werden, das sowohl Messzeitpunkte, Stichproben als auch zu untersuchende Variablen enthielt. Am Beispiel der Unterrichtsassistenten-Evaluation sollen die einzelnen methodischen Überlegungen im Folgenden genauer erläutert werden.

Theoretisches Rahmenmodell: Ableitung von Hypothesen und Zusammenstellung der Messinstrumente

In Abbildung 1 ist dazu das theoretische Rahmenmodell für die Evaluation der Unterrichtsassistenten dargestellt. Darin wurde – auf der Basis theoretischer Überlegungen und empirischer Befunde – angenommen, dass sich die Mitarbeit der Unterrichtsassistenten sowohl auf Lehrkraft- als auch auf Schülerseite auf vielerlei kognitive, affektive und motivationale Variablen auswirken sollte, wobei diese Effekte durch Veränderungen in der Unterrichtsgestaltung und Klassenführung vermittelt werden.

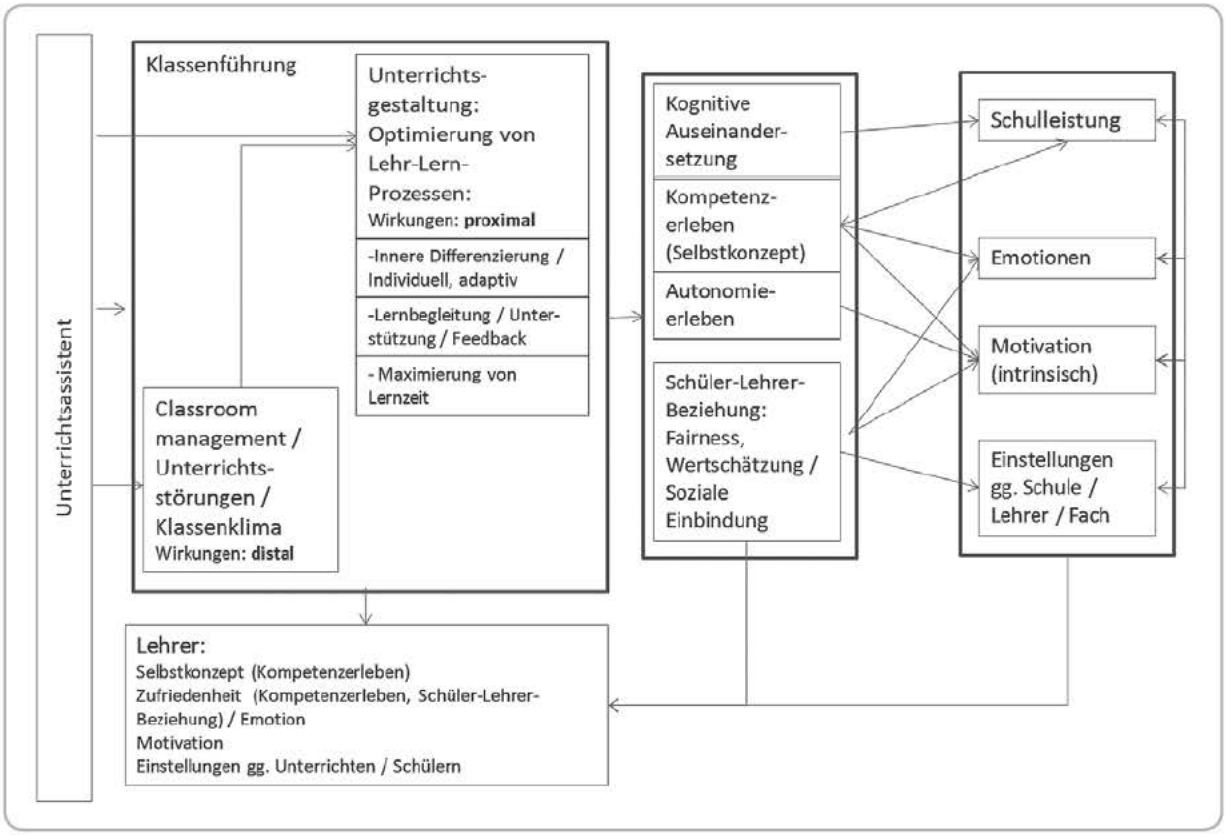

Abb. 1: Theoretisches Rahmenmodell zur Evaluation der Unterrichtsassistenten

Anhand dieses theoretischen Rahmenmodells wurde entschieden, die Perspektiven unterschiedlicher Akteure des Schullebens (Schüler/-innen und Lehrkräfte) in die Evaluation miteinzubeziehen. Außerdem wurden die entsprechenden interessierenden Unterrichts-, kognitiven, affektiven und motivationalen Variablen ausgewählt. Daraus wurden jeweils Schüler/-innen- und Lehrkraftfragebögen zusammengestellt, die zu den zu untersuchenden Variablen wissenschaftlich fundierte Skalen enthielten, die bereits in anderen auch international anerkannten Schulstudien eingesetzt worden waren, wie beispielsweise in den PISA-Studien 2003 bzw. 2006 (PISA-Konsortium 2006; Frey, Taskinen, Schütte & PISA-Konsortium 2009) oder der COACTIV-Studie (Baumert, Blum, Brunner, Dubberke, Jordan, Klusmann et al., 2008).

So wurden sowohl die Schüler/-innen als auch Lehrkräfte zum Beispiel zu Störungen im Unterrichtsgeschehen sowie zur Anwendung von Methoden der internen Differenzierung befragt (als Unterrichtsvariablen); es wurden Lehrkraftselbstwirksamkeit und Schüler/-innen-Selbstkonzept (als kognitive Variablen), Schüler/-innen- und Lehrkraftemotionen (als affektive Variablen) sowie die Motivation sowohl auf Schüler/innen- als auch Lehrkraftseite erfasst. Der Schüler/-innen-Fragebogen bestand schließlich aus mehr als 240 Fragen (Items) in mehr als 50 Skalen, und der Lehrkraftfragebogen aus mehr als 280 Items in über 40 Skalen.

Aus diesen Angaben wird ein typisches Merkmal wissenschaftlicher Befragungen deutlich: Ein Themeninhalt wird üblicherweise in empirischen Untersuchungen nicht nur durch ein einziges Item erfragt, sondern durch mehrere – häufig sehr ähnliche – Fragen abgedeckt, ein Umstand, der von den Befragten als eher unangenehm erlebt und kritisch hinterfragt wird. Tatsächlich bringt jedoch – trotz Einschränkungen in der Testökonomie – die Verwendung mehrerer Items zur Erfassung eines Themengebietes testtheoretisch wertvolle Vorteile mit sich: Zum Einen kann ein einzelnes Item häufig nicht das

Kennwerte der Subskala-Einzelitems

Itemnr.	Itemwortlaut (deutsche Übersetzung): „Wie sicher sind Sie, dass Sie in dieser Klasse …“	Mittelwert	Standardabweichung	Trennschärfe	Cronbachs Alpha, wenn Item weggelassen
1	… Schularbeiten so organisieren können, dass die Anweisungen und Arbeitsaufträge den individuellen Bedürfnissen der Schüler/-innen angepasst sind.	2.89	1.05	.47	.87
2	… allen Schülern/-innen mit unterschiedlichen Leistungsniveaus in dieser Klassen realistische Herausforderungen bieten können.	3.00	.87	.62	.80
3	… Anweisungen an die Bedürfnisse leistungsschwacher Schüler/-innen anpassen können und trotzdem dabei den Bedürfnissen der anderen Schüler/-innen gerecht werden.	3.11	.93	.88	.68
4	… Aufgaben im Unterricht so organisieren können, dass sowohl leistungsstarke als auch leistungsschwache Schüler/-innen an Aufgaben arbeiten, die ihrem Leistungsniveau entsprechen.	2.89	1.05	.71	.76
Kennwerte der zusammengefassten Subskala					
Deskriptive Statistiken		2.97	.80		
Interne Konsistenz (Cronbachs Alpha)					.83

Anmerkungen: Die Antwortskala der Items rangierte von 1 *nicht sehr sicher* bis 5 *sehr sicher*. Die Standardabweichung stellt ein Maß für die Streubreite der in der Stichprobe gemessenen Werte um den Mittelwert dar, d. h. die durchschnittliche Entfernung aller gemessenen Werte vom Mittelwert. Üblichen Empfehlungen zufolge sollten Items mit Trennschärfen < .30 aus der Skala entfernt werden. Für die interne Konsistenz gelten Werte von $\alpha > .70$ als akzeptabel.

Tabelle 1: Item-Trennschärfen, interne Konsistenz (Cronbachs Alpha) sowie deskriptive Statistiken (Mittelwert, Standardabweichung) der Lehrerselbstwirksamkeitssubskala „Adapt Instruction to Individual Needs“ berechnet anhand der Prätest-Lehrkraftstichprobe (Messzeitpunkt t1) des Schuljahres 2011/12

komplette Spektrum eines Konstrukts erfassen, zumal ein Konstrukt sogar mehrere „Subdimensionen“ umfassen kann. Beispielsweise wurde das Konstrukt Lehrerselbstwirksamkeit mit der etablierten Skala von Skaalvik und Skaalvik (2010) erfasst, die aus den vier Subskalen „Instruction“, „Adapt Instruction to Individual Needs“, „Motivate Students“ und „Maintain Discipline“ mit jeweils vier Items besteht.

Zudem verbessert die Verwendung von Multi-Item-Skalen die Reliabilität der Messung, d. h. die Zuverlässigkeit oder Genauigkeit, mit der das Konstrukt erfasst wird. Dabei muss allerdings über statistische Voranalysen sichergestellt werden, dass die einzelnen Items einer Skala untereinander über genügend Gemeinsamkeiten verfügen. Genauer gesagt muss untersucht werden, wie hoch die Korrelation eines einzelnen Items mit den übrigen Items der Skala ausfällt (Trennschärfe); dies bedeutet, wie gut ein Item eine aus den restlichen Items gebildete Skala widerspiegelt, und wie hoch das Ausmaß ist, in dem die Items einer Skala durchschnittlich miteinander korrelieren (interne Konsistenz), d. h. wie eng die Items einer Skala miteinander in Verbindung stehen (vgl. beispielsweise Bühner, 2011). Anhand dieser Kriterien werden Einzelitems zu Multi-Item-Skalen zusammengestellt und auch in den Auswertungen gemeinsam betrachtet, indem aus den Antworten zu den einzelnen Fragen ein Mittelwert für die Gesamtskala gebildet wird. In Tabelle 1 werden die Item-Trennschärfen, die interne Konsistenz sowie deskriptive Statistiken (Mittelwert und Standardabweichung) der Lehrerselbstwirksamkeitssubskala „Adapt Instruction to Individual Needs“, exemplarisch berechnet anhand der Prätest-Lehrkraftstichprobe (Messzeitpunkt t1) des Schuljahres 2011/12, dargestellt.

Prä-Post-Interventions-Kontrollgruppendesign

Desweiteren wurde im Evaluationsdesign festgelegt, zu welchen Messzeitpunkten welche Schüler/-innen- bzw. Lehrkraftstichproben befragt werden sollten, um untersuchen zu können, was sich durch die Einbeziehung von Unterrichtsassistenten ins Unterrichtsgeschehen konkret verändert, welche spezifischen Effekte diese Intervention also bewirken würde. Dieser Frage nach Veränderungen kann wissenschaftlich gesehen am besten durch ein sogenanntes Prä-Post-Design nachgegangen werden. Dies bedeutet, dass die interessierenden Variablen bei den von der Intervention Betroffenen sowohl vor als auch nach Durchführung der Intervention erhoben werden müssen, um Aussagen über mögliche durch sie verursachte Veränderungen machen zu können. Um darüber hinaus ausschließen zu können, dass sich die Variablen lediglich durch die zeitliche Entwicklung quasi „von selbst“ verändern, wurden zusätzlich sogenannte Kontrollgruppen herangezogen, die ebenfalls im Prä- und Posttest befragt

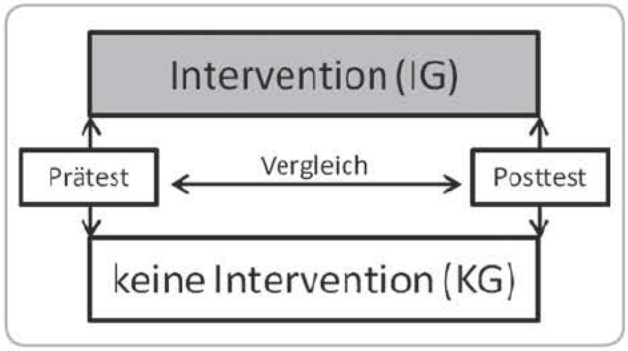

Abb. 2: Vergleichsmöglichkeiten im Prä-Post-Interventions-Kontrollgruppendesign

wurden, ohne jedoch zwischenzeitlich an der Intervention teilzunehmen. Durch Vergleich der Ergebnisse zwischen Interventionsgruppen (IG) und Kontrollgruppen (KG) können eventuell auftretende Effekte auf die Intervention zurückgeführt werden (vgl. Abbildung 2).

Konkret wurde am Gymnasium Wilhelmsdorf ein Design gewählt, das innerhalb jeder zu evaluierenden Jahrgangsstufe (5 – 8) bestimmte Klassen als Interventions- und Kontrollklassen definierte. Während die Interventionsklassen jeweils in ca. sechs Wochen andauernden Phasen an der Intervention in den beiden zur Evaluation ausgewählten Fächern (Deutsch/Mathematik) teilnahmen (d. h. die Deutsch- bzw. Mathematiklehrkräfte kooperierten in dieser Zeit in dieser Klasse mit den Unterrichtsassistenten), wurden die Unterrichtsassistenten in den Kontrollklassen nicht im Unterricht eingesetzt. Die Interventionsphasen in Deutsch bzw. Mathematik fanden nach einem zufälligen Muster in direkt aufeinanderfolgenden sechswöchigen Schul-

	Oster	Pfingst	Sommerferien
Klasse 5a	Mathematik	Deutsch	
Klasse 5b	Deutsch	Mathematik	
Klasse 5c	–	–	
Klasse 6a	Mathematik	Deutsch	
Klasse 6b	–	–	
Klasse 6c	Deutsch	Mathematik	
Klasse 6d	Deutsch	Mathematik	
Klasse 7a	–	–	
Klasse 7b	Mathematik	Deutsch	
Klasse 7c	Mathematik	Deutsch	
Klasse 8a	Deutsch	Mathematik	
Klasse 8b	Mathematik	Deutsch	
Klasse 8c	–	–	

Abb. 3: Übersicht über den Interventionsplan im Schuljahr 2011/12. In den Klassen 5c, 6b, 7a und 8c fand in diesen Zeiträumen keine Intervention statt, so dass es sich hierbei um die „Kontrollklassen“ handelt.

	Schuljahr 2011/12			Schuljahr 2012/13				
	t1	t2	t3	t4	t5	t6	t7	t8
Lehrkräfte mit Intervention	0	7	8	0	2	7	5	2
Lehrkräfte ohne Intervention	9	5	16	12	9	8	6	2
Lehrkräfte gesamt	9	12	24	12	11	15	11	4
Schüler/-innen mit Deutsch-Intervention	0	106	64	0	0	108	42	23
Schüler/-innen mit Mathematik-Intervention	0	88	109	0	69	41	62	15
Schüler/-innen ohne Intervention	278	86	82	161	88	23	22	0
Schüler/-innen gesamt	278	280	255	161	157	172	126	38

Tabelle 2: Übersicht über die Stichprobenumfänge in den Schuljahren 2011/12 und 2012/13

jahresphasen statt. Exemplarisch findet sich eine Übersicht über den ersten Interventionsplan nach Projektstart in der zweiten Hälfte des Schuljahres 2011/12 in Abbildung 3. Jeweils zu Beginn und am Ende jedes Zeitintervalls wurden die Schüler/-innen sowie die Deutsch- und Mathematik-Lehrkräfte sowohl der Interventions- als auch der Kontrollklassen mittels Fragebögen befragt.

Zwischenstand und Ausblick

In den Schuljahren 2011/12 und 2012/13 wurden für die Evaluation der Unterrichtsassistenten bislang insgesamt N = 1467 Schüler/-innen- und N = 98 Lehrkräftefragebögen aus 17 Klassen ausgefüllt. Details über die Stichprobenumfänge in den einzelnen Befragungsbedingungen sowie zu den verschiedenen Messzeitpunkten sind Tabelle 2 zu entnehmen. Darüber hinaus wurden alle Mitglieder des Lehrerkollegiums (N = 48) vor Projektbeginn zu den Osterferien des Schuljahres 2011/12 über ihr aktuelles Befinden sowie ihre Erwartungen an die Kooperation mit den Unterrichtsassistenten befragt.

Das Untersuchungsdesign wird im derzeit laufenden Schuljahr 2013/14 sowie voraussichtlich auch im Schuljahr 2014/15 mit den Klassen der jeweils neu beginnenden 5. Jahrgangsstufen fortgeführt. Außerdem umfasst die wissenschaftliche Begleitung seit Beginn des Schuljahres 2013/14 auch die zweite neue personelle Säule des „Power of Difference"-Projekts, nämlich die Mitarbeit der Förderpädagogen bzw. Coaches. Bei der Coach-Evaluationsplanung wurde prinzipiell genauso vorgegangen wie bei der Evaluation der Unterrichtsassistenten beschrieben. Jedoch liegt dort aufgrund der Art der Intervention, die – im Gegensatz zum Aufgabengebiet der Unterrichtsassistenten – eher eine individual- als klassenbasierte Unterstützung vorsieht, ein stärkerer Fokus auf den einzelnen Schüler/-innen, die die Beratung bzw. Förderung durch den Coach in Anspruch nehmen.

Durch die exakte methodische Vorgehensweise bei der Durchführung und Evaluation von „Power of Difference" sind aus wissenschaftlicher Perspektive die strengen Voraussetzungen dafür erfüllt, dass die empirischen Evaluationsergebnisse dieses Projekts zu zulässigen und interpretierbaren Befunden führen, die ein solides Fundament für evidenzbasierte bildungspolitische Diskussionen und Entscheidungen liefern.

Literatur

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., Kunter, M., Löwen, K., Neubrand, M., & Ts, Y.-M.: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente [Materialien aus der Bildungsforschung] (Vol. 83). Berlin 2008. Max-Planck-Institut für Bildungsforschung.

Bühner, M.: Einführung in die Test- und Fragebogenkonstruktion. München 2010.

Frey, A., Taskinen, P., Schütte, K., & PISA-Konsortium Deutschland (Hrsg.): PISA 2006 Skalenhandbuch – Dokumentation der Erhebungsinstrumente. Münster 2009.

Hager, W.: Evaluation pädagogisch-psychologischer Interventionsmaßnahmen. In: W. Schneider & M. Hasselhorn (Hrsg.): Handbuch der Pädagogischen Psychologie (S. 721-732) (Handbuch der Psychologie, Bd. 10). Göttingen 2008.

PISA-Konsortium Deutschland (Hrsg): PISA 2003: Dokumentation der Erhebungsinstrumente. Münster 2006.

Skaalvik, E. M., & Skaalvik, S.: Teacher Self-Efficacy and Teacher Burnout: A Study of Relations. Teaching and Teacher Education (2010), 26, 1059-1069.

Wild, E. & Möller, J.: Pädagogische Psychologie. Heidelberg 2009.

Dr. Birgit Mirjam Beisswingert
birgit.beisswingert@uni-konstanz.de

Prof. Dr. Thomas Götz
thomas.goetz@uni-konstanz.de

Universität Konstanz, Pädagogische Hochschule Thurgau