

TIPP – erster Zwischenbericht der Bielefelder Evaluation

Erstes Modul Wahrnehmung, Kommunikation, Konfliktmanagement

Prof. Dr. Rainer Dollase und Dipl.-Päd. Odette Selders
Universität Bielefeld, Abt. Psychologie und IKG

Januar 2009

1. Fragestellung

Die Notwendigkeit einer empirischen Evaluation pädagogischer Programme und Projekte ist heute unbestritten. Zu oft hat sich in der Vergangenheit gezeigt, dass gut gemeinte Ansätze ihre behaupteten Wirkungen nicht erreichen. Daraus hat sich eine gewisse Skepsis gegenüber zwar gut begründeten und gemeinten Programmen entwickelt, die zu der zunehmend häufigeren empirischen, meist auch quantitativen, Evaluation der Effekte geführt hat.

Die Effektmessung als solche ist ein umfängliches wissenschaftliches Gebiet und keineswegs unproblematisch. Selten sind die Verfahren, die eingesetzt werden, so sensibel, dass sie auch kleinste Veränderungen nachzeichnen können, und andererseits ist man immer wieder im so genannten „Bandbreite-Fidelitäts-Dilemma“ gefangen, d.h. eine ausführliche und umfangreiche Befragung wäre eigentlich günstig, würde sich aber spezialisieren müssen und die Breite möglicher Effekte nicht erfassen können. Solche und ähnliche Probleme sind seit Jahrzehnten bekannt und stellen sich auch in dieser Untersuchung.

Von besonderer Bedeutung ist bei einer Evaluation einmal die Längsschnittlichkeit der Untersuchung und zum anderen die Existenz einer Kontrollgruppe. Wirkungsevaluation kann nur über Längsschnitte erfolgen, da man nachweisen muss, dass sich die Parameter im Laufe des Trainings oder des Programms gebessert haben. Eine Kontrollgruppe ist notwendig, damit der Vorteil des Programms oder Trainings gegenüber den herkömmlichen bzw. gegenüber anderen Ansätzen erwiesen werden kann.

Im TIPP-Projekt heißt das, dass folgende Fragestellungen beantwortet werden müssen:

1. Welche längsschnittlichen Effekte zeigt das Trainingsprogramm der Bielefelder Gruppe des TIPP-Projektes?
2. Wie unterscheidet sich die Bielefelder Gruppe von einer Kontrollgruppe?

Für den ersten Zwischenbereich der Evaluation ist insbesondere interessant, ob die Ausgangslagen zwischen Trainingsgruppe und Kontrollgruppe vergleichbar waren, denn wenn Ausgangsunterschiede existieren, müssten diese durch eine so genannte Kovarianzanalyse (Ausgangslage als Kovariat) kontrolliert werden.

Das Bandbreite-Fidelitäts-Dilemma wird durch einen kurzen, d.h. nur zweiseitigen Fragebogen gelöst, d.h. die Fragen und Items sind auf globale Einschätzungen der 3

inhaltlich wichtigen Bereiche „Kommunikation“, „Konfliktbearbeitung“ und „Klassenführung“ bezogen. Genau diese drei Kompetenzbereiche werden in der Bielefelder Trainingsgruppe angeboten. Die Aufnahme der drei Bereiche im Fragebogen erfüllt die Forderung nach „curricularer Validität“ eines Evaluationsfragebogens, d.h. er soll das erfragen bzw. messen, was auch Anlass und Ziel des Trainings ist.

2. Stichprobe und Methode

Wie in der Abb. 1 erkenntlich, haben insgesamt 103 Referendare an der Befragung teilgenommen. Diese 103 Referendare und Referendarinnen teilen sich auf in 56 Personen in der Versuchsgruppe und 47 in der Kontrollgruppe. In der Versuchsgruppe haben nur 22 an dem Training teilgenommen, in der Kontrollgruppe logischerweise keiner.

Abb. 1: Stichprobengrößen von Trainings- und Kontrollgruppe

	Training ja	Training nein	Gesamt
Versuchsgruppe	22	34	56
Kontrollgruppe	0	47	47
Gesamt	22	81	103

Wie Abb. 2 zeigt, haben nur 18 Personen an allen 3 Erhebungen teilgenommen (t1, t2 und t3), bei der dritten Erhebung sogar einige Personen mehr (n=22). Die erste Erhebung fand kurz nach dem Eintritt in das Referendariat statt, die zweite Erhebung und die dritte Erhebung innerhalb von 8 Wochen nach Beginn des Referendariats.

Abb. 2: Trainingsgruppe und Kontrollgruppe , Erhebungszeitpunkte, Stichproben

	Trainingsgruppe	Kontrollgruppe
1.Erhebung	18	80
2.Erhebung	19	33
3.Erhebung	22	0

Der Fragebogen enthielt neben der Anrede der Referendare und Referendarinnen und der Garantie einer anonymen Beantwortung und Auswertung eine Codierung, die notwendig wurde, weil es sich ja um eine anonyme Längsschnittstudie handelt. Als Code wurde der Geburtsmonat der Mutter, der zweite Buchstabe im Vornamen des Vaters und der letzte Buchstabe des Geburtsortes verlangt. Dadurch war es möglich, in allen Fällen die Fragebögen zuzuordnen.

Im *ersten Fragebogenblock* waren neun Items enthalten, die jeweils für Kommunikation, Konfliktbearbeitung und Klassenführung drei Beurteilungen verlangten und zwar sollte die „Güte der Ausbildung“, das „heutige Wissen“ und das „jetzige Verhalten“ in den drei Bereichen mit einer Schulnote von 1 bis 6 beurteilt werden. Dadurch wird es möglich, subjektive Einschätzungen von Ausbildung, Wissen und Verhalten über die drei Erhebungszeitpunkte zu erfassen.

Im *zweiten Block* wurden Noten für die Wichtigkeit der drei Bereiche Kommunikation, Konfliktbearbeitung und Klassenführung erfragt (3 Items).

Im *dritten Block* wurde ein vollständiger Paarvergleich (sog. pair comparison technique) verlangt. Und zwar sollte das „Fachwissen in den Fächern“, die „Verbesserung der Kommunikation“, die „Verbesserung der Konfliktbearbeitung“, die „Verbesserung der Klassenführung“, die „Unterrichtsabläufe in den Fächern“ in allen Zweierkombinationen beurteilt werden. Die Frage lautete: „Entscheiden Sie sich jeweils für eine von zwei Alternativen, was ist für die Praxis wichtiger?“ Dann folgten alle möglichen Zweierkombinationen, z.B. Verbesserung meiner Konfliktbearbeitung oder Verbesserung meiner Klassenführung. Mit Hilfe des Paarvergleichsverfahrens kann man besonders zuverlässige Rangreihen der Wichtigkeit erfassen.

Im *vierten Block* wurden die Erwartungen der Referendare bezüglich der größten Herausforderungen in der Praxis erfragt. Die Referendare und Referendarinnen sollten die Höhe der Schwierigkeitserwartung in Schulnoten von 1 = geringe Schwierigkeiten bis 6 = allergrößte Schwierigkeiten angeben. Hier wurden insgesamt 6 (in der zweiten und dritten Messung 7) Themen genannt. Beispiele: Den Schülern/Schülerinnen den Stoff verständlich vermitteln können. Allen Schülern gerecht zu werden. Der Umgang mit Heterogenität der Schülerschaft.

Zum Schluss wurde noch nach Geschlecht und Alter gefragt, auch nach der fachwissenschaftlichen Orientierung, ob sich der Referendar/die Referendarin eher naturwissenschaftlich, eher gesellschaftswissenschaftlich oder eher sprachwissenschaftlich orientiert.

Der Fragebogen ist im Anhang dieses Berichtes enthalten.

3. Ergebnisse

3.1. Die Prüfung von Anfangsunterschieden zwischen Kontrollgruppe und Trainingsgruppe

Für den Nachweis von Effekten des Trainings ist die Prüfung der Ausgangsunterschiede zwischen Trainings- und Kontrollgruppe wichtig. In Abb. 3 sind die entsprechenden Mittelwerte bei Versuchs- und Kontrollgruppe im ersten Messzeitpunkt aufgeführt. Es gibt nur zwei signifikante Unterschiede, und zwar im „Verhalten Kommunikation“, hier ist die Kontrollgruppe besser, und im „Verhalten Klassenführung“, auch hier fühlt sich die Kontrollgruppe etwas besser. Der erste Befund ist auf dem 2%-Niveau, der zweite auf dem 5%-Niveau signifikant.

Abb.3: Beurteilung von Ausbildung, Wissen und Verhalten. Vergleich der Versuchsgruppe mit der Kontrollgruppe zum 1. Messzeitpunkt. * = Signifikanter Unterschied

	Versuchsgruppe	Kontrollgruppe
Ausbildung Kommunikation	3,7	3,3
Ausbildung Konfliktbearbeitung	4,4	4,1
Ausbildung Klassenführung	4,5	4,4
Wissen Kommunikation	3,1	2,9
Wissen Konfliktbearbeitung	3,6	3,3
Wissen Klassenführung	4,0	3,8
Verhalten Kommunikation	3,0*	2,6*
Verhalten Konfliktbearbeitung	3,2	2,9
Verhalten Klassenführung	3,8*	3,4*

Um eine multivariate Prüfung der Anfangsunterschiede durchzuführen wurde eine Diskriminanzanalyse gerechnet. Hierbei wurden alle Ausbildungs-, Wissens- und Verhaltensfragen des Fragebogenblocks als abhängige Variablen in die Diskriminanzanalyse zur Unterscheidung von Gruppe 1 (Trainings-/Versuchsgruppe) und Gruppe 2 (Kontrollgruppe) eingegeben. Das Klassifikationsresultat ist nicht sehr deutlich. 61% der Fälle, also nur 11% mehr als Zufall, werden korrekt klassifiziert.

In Abb. 4 findet ein Vergleich von Trainings- und Kontrollgruppe in der ersten Messung zum Fragebogenblock der „Herausforderung der Praxis“ statt. Hier sind alle Items nicht signifikant, d.h. es gibt keine Unterschiede zwischen Versuchs- und Kontrollgruppe. Eine entsprechende Diskriminanzanalyse ermittelt demzufolge auch nur 57% korrekt klassifizierte Fälle, d.h. 7% mehr als Zufall.

Die absoluten Notenwerte sind interessant: „den richtigen Ton treffen“ wird als nicht so schwierig wie z.B. die Forderungen „Schüler ruhig halten“ oder „Allen gerecht werden“ beurteilt.

Abb. 4: Erwartete größte Herausforderung in der Praxis – von 1=geringe bis 6 = größte Schwierigkeiten. Unterschiede zwischen Versuchs- und Kontrollgruppe in der 1. Messung sind sämtlich nicht signifikant

	Versuchsgruppe(N=56)	Kontrollgruppe(N=47)
Schüler ruhig halten	3,4	3,4
Fachwissen haben	3,0	2,7
richtigen Ton treffen	2,6	2,4
Konflikte effektiv lösen	3,2	3,3
Stoff verständlich machen	3,0	2,8
Allen gerecht werden	3,7	3,9

3.2 Längsschnittliche Unterschiede in der Trainingsgruppe

Die Erwartungen des TIPP - Projektes in Bielefeld sind, dass sich über drei Messzeitpunkte Verbesserungen in der Einschätzung von Ausbildung, Wissen und Verhalten in den drei zentralen Inhaltsbereichen ergeben. Hiermit kovariierend sollte die Beurteilung der Herausforderung unterschiedlicher Praxisprobleme sinken und auch die Wichtigkeitsbeurteilung einer Verbesserung ebenfalls sinken.

In Abb. 5 sind die Ergebnisse über die Zeitpunkte der Trainingsgruppe in Ausbildung, Wissen und Verhalten dargestellt. T1 wird gegen t3 geprüft und man erkennt, dass die Ausbildung in Kommunikation, Konfliktbearbeitung und Klassenführung deutlich besser geworden ist (signifikant). Ebenfalls gilt dies für das „jetzige Wissen“ in Kommunikation und Konfliktbearbeitung, auch das wird bei t3 signifikant besser beurteilt, nicht aber das Wissen im Bereich Klassenführung, hier ergibt sich ein nicht signifikantes Ergebnis.

Abb. 5: Ausbildung, Wissen und Verhalten. Entwicklung in der Trainingsgruppe über drei Zeitpunkte . *= signifikante Veränderung von t1 nach t3; nsf = nicht signifikante Veränderung von t1 nach t3

	t1	t2	t3
Ausbildung Kommunikation	3,6	3,2	2,4*
Ausbildung Konfliktbearbeitung	4,1	3,8	2,5*
Ausbildung Klassenführung	4,4	3,5	3,5*
Wissen Kommunikation	2,9	2,7	2,2*
Wissen Konfliktbearbeitung	3,1	3,1	2,6*
Wissen Klassenführung	3,6	3,2	3,4nsf
Verhalten Kommunikation	2,8	2,6	2,6nsf
Verhalten Konfliktbearbeitung	2,9	2,9	2,8nsf
Verhalten Klassenführung	3,5	3,4	3,1nsf

Besonders überraschend und interessant ist, dass das *Verhalten* sowohl in Kommunikation als auch Konfliktbearbeitung und Klassenführung sich nicht signifikant verbessert. D.h. die Referendare nehmen einen Ausbildungs- und Wissensgewinn von dem bisherigen Training mit, nicht aber spüren sie, dass sich ihr tatsächliches Verhalten in diesen drei Bereichen signifikant gebessert hätte.

Abb. 6 zeigt, dass die Beurteilung der Herausforderung unterschiedlicher Anforderungen des Schulalltages, zum Beispiel „Konflikte effektiv lösen“, von t1 nach t3 sinkt. D.h. es gibt eine (in allen Fällen nicht signifikante) Tendenz zur Beantwortung in Richtung geringere Schwierigkeiten bei t3.

Abb.6: Erwartung praktischer Schwierigkeiten in der Trainingsgruppe.

	t1	t2	t3
Schüler ruhig halten	3,5	3,2	2,8
Fachwissen haben	2,8	2,7	2,5
richtigen Ton treffen	2,4	2,2	2,3
Konflikte effektiv lösen	3,1	3,3	2,8
Stoff verständlich machen	2,9	2,6	2,4
Allen gerecht werden	3,7	3,6	3,4
Heterogenität	-	3,6	3,4

Ebenfalls erwartungskonform ist, dass die Wichtigkeitsbeurteilung der Verbesserung von Kommunikation, Konfliktbearbeitung und Klassenführung über die drei Zeitpunkte sinkt – zwar nicht viel, aber doch in allen drei Bereichen synchron.

Abb. 7: Wichtigkeitsbeurteilung in der Trainingsgruppe über die drei Zeitpunkte

	t1	t2	t3
Wichtigkeit Verbesserung Kommunikation	1,4	1,6	1,7
Wichtigkeit Verbesserung Konfliktbearbeitung	1,3	1,6	1,7
Wichtigkeit Verbesserung Klassenführung	1,3	1,6	1,8

Der *systematische Paarvergleich* der Praxiswichtigkeit von Fachwissen, Unterrichtsabläufen, Verbesserung von Kommunikation, Konfliktbearbeitung und Klassenführung zeigt über die drei Zeitpunkte eine geringfügige Veränderung. In allen drei Zeitpunkten bleiben die letzten beiden Plätze konstant, nämlich Unterrichtsabläufe in meinen Fächern (Rangplatz 3) und Fachwissen (Rangplatz 4). Bei t1 und t2 liegen Kommunikation- und Konfliktbearbeitung gemeinsam auf Rang 2 und Klassenführung auf Rang 1. Beim Zeitpunkt t3 gibt es an der Spitze einen Wechsel: Kommunikation und Konfliktbearbeitung liegen auf Platz 1 und Klassenführung auf Platz 2.

Es wurde zum Zeitpunkt t3 geprüft, ob Geschlechtsunterschiede, Altersunterschiede und Unterschiede gemäß der fachlichen Orientierung geisteswissenschaftlich, sprachwissenschaftlich oder gesellschaftswissenschaftlich existieren. Hier konnten keine signifikanten Zusammenhänge gefunden werden.

Abb. 8: Systematischer Paarvergleich - Prozentwerte und Rangplätze (in Klammern)

	Fachwissen	Unterrichts abläufe	Kommun ikation	Konflikte	Klassen- führung
T1	7 (4)	16 (3)	21 (2)	21 (2)	27 (1)
T2	12 (4)	19 (3)	21 (2)	21 (2)	26 (1)
T3	11 (4)	18 (3)	25 (1)	25 (1)	20 (2)

4. Diskussion

Zunächst einmal muss relativierend angemerkt werden, dass in der Trainingsgruppe längsschnittlich nur eine relativ geringe Anzahl von Referendaren enthalten war (N=18 bis N=22). Das hat statistisch den Effekt, dass nur erhebliche Unterschiede signifikant werden. Auch ist die Signifikanzprüfung bei kleinen Stichproben nicht unbedingt immer mit Relevanz zu übersetzen. Relevant können auch kleinere und nicht signifikante Verbesserungen sein.

Zunächst konnte aber gesichert werden, dass die Anfangsunterschiede zwischen Trainingsgruppe und Kontrollgruppe nicht signifikant voneinander abweichen. Bis auf zwei Items im Bereich der Beurteilung von Kommunikations- und Klassenführungsverhalten gibt es keine signifikanten bzw. relevanten Unterschiede zwischen Trainings- und Kontrollgruppe. Das gilt auch für die Beurteilung der praktischen Herausforderungen unterschiedlicher Lehraufgaben, die alle nicht signifikant sind. Mit den entsprechenden Diskriminanzanalysen konnte dieser Befund erhärtet werden. Es ist aus unserer Sicht zu beurteilen: Im Wesentlichen sind sich Trainings- und Kontrollgruppe zu Beginn des Experimentes vergleichbar. Das ist eine Grundvoraussetzung für die Ermittlung von tatsächlichen Effekten.

Die längsschnittlichen Ergebnisse in der Trainingsgruppe zeigen zunächst einmal einen allgemeinen interessanten Befund, nämlich dass die Ausbildungsbeurteilung und das Wissen sich tatsächlich im Laufe des Trainings in den drei Bereichen Kommunikation, Konfliktbearbeitung und Klassenführung verbessert. Aber: die

Beurteilung des eigenen Verhaltens verbessert sich nicht signifikant und auch nicht praktische relevant. Hier ist also ein Defizit bzw. es muss konstatiert werden, dass eine gute Ausbildung und ein gutes Wissen über diese Bereiche offenbar nicht reicht, um das Verhalten tief greifend zu verbessern.

Dass die Wichtigkeit der Verbesserung von Kommunikation, Konfliktbearbeitung und Klassenführung sinkt, ist auf den zunehmenden Kompetenzgewinn der Referendare und Referendarinnen zurück zu führen, also ein durchaus positiver Befund. Das gilt auch über das geringer werden des erwarteten Schwierigkeitsgrades für: „die Schülerinnen im Unterricht ruhig und diszipliniert zu halten“, „dass nötige Fachwissen sicher zu beherrschen“ etc. – das tendenzielle Sinken der Schwierigkeitserwartungen von t1 zu t3 ist also positiv zu werten.

Der Paarvergleich zeigt, dass zwischen t2 und t3 in der praktischen Relevanzbeurteilung ein kleiner Hierarchiewechsel eingetreten ist. Kommunikation und Konfliktbearbeitung wird für wichtiger gehalten als die Beherrschung der Klassenführungstechniken. Das kann der Tatsache geschuldet sein, dass Klassenführungsverhalten natürlich sowohl Kommunikation als auch Konfliktbearbeitung ist, bzw. dass Kommunikation und Konfliktbearbeitung wesentliche Bestandteile des Klassenführungsverhaltens sind. Möglicherweise müsste aber die Ausbildung in Klassenführungstechniken verbessert werden.

17.1.2009