

Qualitätskontrolle von Qualitätsbeurteilungen - Die Analyse von Expertenübereinstimmungen

Siegfried Preiser & Sonja Wermuth

Anmerkung: Dieser Beitrag basiert auf einem Vortrag auf dem Deutschen Psychologentag 2005 / Kongress für Angewandte Psychologie, 10.-12. November 2005 in Potsdam (Preiser & Wermuth, 2005).

Die Sektion Politische Psychologie hat in Zusammenarbeit mit einem Expertenbeirat einen Katalog von Qualitätskriterien zur Beurteilung von Gewaltpräventionsprogrammen in Form einer Checkliste entwickelt (Preiser & Wagner, 2003; vgl. auch den Beitrag „Qualitätskriterien für Präventionsprogramme gegen Gewalt, Rechtsextremismus und Fremdenfeindlichkeit: Ein Projekt der Sektion Politische Psychologie“ von Siegfried Preiser in diesem Band). Diese Checkliste enthält insgesamt 56 Einzelkriterien zu sieben Bereichen: Zielklärung, Zielgruppe, Theoretische Grundlagen, Maßnahmenbeschreibung, Kompetenzen der Trainerinnen und Trainer, Evaluation und Qualitätssicherung sowie Preis-Leistungs-Verhältnis. Dieser Kriterienkatalog soll fortlaufend dem aktuellen fachlich-wissenschaftlichen Diskussionsstand angepasst und hinsichtlich seiner Praktikabilität überprüft werden. Einige Ergebnisse hierzu wurden bereits auf der 16. Bundeskonferenz für Schulpsychologie 2004 in Nürnberg (Preiser, 2006) vorgestellt.

Beurteilerübereinstimmungen in Form von Profilkorrelationen

Um Beurteilerübereinstimmungen quantifizieren zu können, wurden die 56 Qualitätskriterien der Checkliste zu 31 Aspekten zusammengefasst, die jeweils auf einer 4-stufigen Schätzskala (1 = Kriterium kaum erfüllt bis 4 = Kriterium voll bzw. optimal erfüllt) zu beurteilen sind. In drei Universitätsseminaren wurden jeweils von ein oder zwei Referenten bzw. Referentinnen insgesamt 16 verschiedene Präventionsprogramme vorgestellt, davon sechs mehrfach in verschiedenen Seminaren. Die Referenten beurteilten unabhängig voneinander das Programm, mit dem sie sich intensiv auseinander gesetzt hatten. Nach der Präsentation und Diskussion erfolgte die Beurteilung des Programms durch den Seminarleiter sowie die Zuhörerinnen und Zuhörer. Die Urteile der Studierenden wurden gemittelt, und zwar getrennt für Hauptfachstudierende im Diplomstudiengang Psychologie und Nebenfachstudierende. Die Beurteilerübereinstimmung wurde für jedes Programm in Form einer Ähnlichkeitsmatrix auf

der Basis von Profilkorrelationen berechnet. Korrelationen zwischen verschiedenen Beurteilern drücken bei dieser Art der Auswertung aus, wie hoch die Übereinstimmungen bezüglich der Stärken und Schwächen des jeweiligen Programms sind.

hier etwa Abb 1 und 2

Die bisherigen Befunde lassen sich wie folgt zusammenfassen und interpretieren (vgl. Abbildung 1 und 2; Preiser, 2006):

1. Zwei Referenten als Quasi-Experten, die das gleiche Präventions- oder Interventionsprogramm bearbeiten, erreichen im Regelfall mittlere Übereinstimmungswerte sowohl untereinander als auch im Vergleich zu den Programmbewertungen des Seminarleiters und der Zuhörer.
2. Obwohl das Beurteilungssystem allen Beteiligten bekannt ist, weisen die Übereinstimmungswerte große Unterschiede auf (von -.21 bis .88). Bei Nebenfachstudierenden als Referenten oder als Zuhörergruppe gab es – im Vergleich zu Hauptfachstudierenden – etwas häufiger Ausreißer in Richtung niedriger Übereinstimmungswerte, vereinzelt sogar mit negativen Korrelationskoeffizienten.
3. Einzelne Programme sind – auch aus Sicht des Seminarleiters – von der Substanz bzw. den vorliegenden Veröffentlichungen her wenig eindeutig bezüglich ihrer Qualitätskriterien. Hier ergaben sich auch die größten Beurteilungsunterschiede, v.a. zwischen Hauptfach- und Nebenfachstudierenden.
4. Es mag überraschend erscheinen, dass die Übereinstimmungen zwischen Seminarleiter und Referenten geringer sind und eine größere Bandbreite aufweisen als die Übereinstimmungen zwischen Seminarleiter und Zuhörern sowie zwischen Referenten und Zuhörern. Hierbei handelt es sich vermutlich um ein reines Methodenartefakt: Bei den Urteilen der Zuhörer handelt es sich um Mittelwerte mehrerer Personen, weshalb die Fehlervarianz reduziert und die Reliabilität erhöht ist.

Beurteilerübereinstimmungen in Form von Intraklassenkorrelationen

In einem weiteren Auswertungsschritt und unter Berücksichtigung zusätzlicher Daten wurden Intraklassenkorrelationen (ICC – vgl. Domsch & Gerpott, 1986) berechnet. Der Kennwert ICC (1) drückt aus, wie weit sich die Beurteiler jeweils bezüglich der Programme einig sind, wie hoch also die Objektivität der Einzelurteile ist. Der Kennwert ICC (2) drückt aus, wie hoch die Reliabilität der über mehrere Beurteiler zusammengefassten Urteile ist, inwieweit also die Beurteilungen insgesamt eine zuverlässige Differenzierung zwischen den

Programmen ermöglichen. Die ICC-Werte errechnen sich varianzanalytisch über das Verhältnis der Varianz innerhalb der Beurteilungen eines Objektes und der Varianz zwischen den Programmen nach folgenden Formeln:

$$\text{ICC (1)} = \frac{MQ_z - MQ_i}{MQ_z + (n-1) MQ_i} \qquad \text{ICC (2)} = 1 - \frac{MQ_i}{MQ_z}$$

MQ_z = Mittlere quadratische Abweichung der Kollektivmittelwerte vom Gesamtmittelwert

MQ_i = Mittlere quadratische Abweichung der Einzelmesswerte vom Mittelwert des jeweiligen Kollektivs

n = Zahl der Befragten (hier: Gutachter) pro Kollektiv.

Für kleine Stichproben und ungleiche Befragtenzahlen wurden noch statistische Korrekturen angebracht (vgl. Domsch & Gerpott, 1986; Giesler, 2003).

ICC-Werte werden nicht – wie die oben dargestellten Profilkorrelationen – für einzelne Beurteilungsobjekte über alle Kriterien berechnet, sondern für einzelne Kriterien über alle Objekte. Der ICC (2) ist in der Regel höher als der ICC (1). Selbst wenn die Übereinstimmung der einzelnen Beurteiler gering sein sollte, können die gemittelten Werte mehrerer Beurteiler eine zuverlässige Differenzierung ermöglichen, weil sich die individuellen Fehleranteile ausgleichen.

ICC-Werte lassen sich für alle einzelnen Beurteilungskriterien berechnen. So kann man zeigen, welche Kriterien zuverlässig durch die Beurteiler- bzw. Gutachtergruppe beurteilt und zur Unterscheidung der verschiedenen Programme herangezogen werden können. Es ist allerdings zu berücksichtigen, dass hohe Übereinstimmungswerte nahe 1,00 nur möglich sind, wenn alle folgenden Bedingungen optimal erfüllt sind:

1. Die zu beurteilenden Kriterien müssen eindeutig sein und von allen Beurteilern gleichsinnig interpretiert werden.
2. Die Beurteiler müssen fachlich kompetent sein, um die verfügbaren Informationen zu einem Urteil verarbeiten zu können.
3. Die für eine zuverlässige und gültige Beurteilung erforderlichen Informationen müssen zur Verfügung stehen.
4. Die zu beurteilenden Objekte (hier: Präventionsprogramme) müssen sich hinsichtlich der zu beurteilenden Dimensionen deutlich unterscheiden.

Wir haben uns bemüht, für die erstgenannte Bedingung mit unserem Kriterienkatalog optimale Voraussetzungen zu schaffen. Bei den anderen Bedingungen gibt es jedoch Einschränkungen, die hohe Übereinstimmungswerte verhindern: (Zu 2.) In unserer Studie waren die „Gutachter“ - neben dem Seminar- und Projektleiter - Studierende der Pädagogik und der Psychologie, deren fachliche Expertise noch nicht vorausgesetzt werden kann. (Zu 3.) Die publizierten, den Referenten zur Verfügung stehenden Programmbeschreibungen enthielten zu manchen Kriterien nur spärliche Informationen. (Zu 4.) In den Seminaren wurden fast ausschließlich bekannte und bewährte Programme vorgestellt, so dass für die meisten Kriterien nur eine eingeschränkte Varianz zwischen den Programmen vorlag; die Varianz ist aber gerade die Basis für die Ermittlung zuverlässiger Beurteilungsunterschiede. Angesichts dieser Einschränkungen können ICC (2)-Werte ab .50 als brauchbar betrachtet werden, weil unter Ernstbedingungen, d.h. bei ausgewiesenen Experten, denen umfassendes Material über die zu begutachtenden Programme vorliegt, mit demselben Beurteilungssystem deutlich höhere Werte zu erwarten sind.

ICC(2)-Werte wurden für alle 31 Beurteilungsdimensionen (Itemebene) auf der Basis von 22 beurteilten Programmen (davon 6 doppelt) berechnet. Außerdem wurden die Teildimensionen für die sieben Hauptbereiche des Beurteilungsbogens zusammengefasst und gemittelt. Für diese sieben gemittelten Urteile wurden sowohl ICC (1) als auch ICC (2)-Werte berechnet. Dabei flossen nur die Beurteilungen der Quasi-Experten (Referenten und Seminarleiter) in die Berechnungen ein.

Ergebnisse

In der als Anlage beigefügten Tabelle 1 werden für alle quantitativ zu beurteilenden Kriterien ICC (2)-Werte auf der Basis von 22 beurteilten Programmen (davon 6 doppelt) mit jeweils 2 bis 3 Quasi-Gutachtern (Referenten und Seminarleiter) dargestellt, außerdem Mittelwerte und Standardabweichungen der Urteile auf jeweils 4-stufigen Ratingskalen, weiterhin aus der zugrunde liegenden Varianzanalyse die F- und p-Werte. Signifikante F-Werte drücken aus, dass sich die beurteilten Programme – aus der Sicht der Beurteilergruppe – hinsichtlich des jeweiligen Kriteriums bedeutsam voneinander unterscheiden lassen. Zusätzlich werden ICC (1)- und ICC (2)-Werte für die sieben Hauptkriterien, die aus jeweils 2 bis 7 Einzelkriterien gemittelt wurden, dargestellt.

Es zeigt sich, dass manche Kriterien recht reliabel beurteilt werden und deshalb eine klare Differenzierung zwischen verschiedenen Programmen erlauben, beispielsweise

- Wird deutlich, anhand welcher nachprüfbarer Kriterien der Erfolg der Maßnahme überprüft werden kann?
- Werden die einzelnen Maßnahmen aus den theoretischen Grundlagen abgeleitet?
- Gibt es Aussagen zu den Rahmenbedingungen?
- Welche Methoden kommen zum Einsatz? Wie werden deren erwartete Wirkungen begründet?
- Wie werden die Anwender, Mediatoren oder Multiplikatoren des Programms ausgebildet, eingewiesen und supervidiert?
- Werden vergleichbare Kontrollgruppen oder Wartekontrollgruppen berücksichtigt?

Andere Kriterien werden überhaupt nicht reliabel beurteilt; teilweise ist die Varianz innerhalb der Gruppen größer als die Varianz zwischen den Gruppen, was zu negativen ICC-Werten führt, beispielsweise

- Gibt es Aussagen über die Interventionsziele?
- Werden realistische Effekte erwartet und quantifiziert bzw. präzisiert?
- Welche Schritte zur Qualitätssicherung des Programms sind geplant?
- Welche Effekte werden erwartet, in welcher Höhe?
- Werden Langzeiteffekte, Multiplikatoreffekte und positive Nebenwirkungen erwartet?

Insbesondere Fragen zur Evaluation und zum Preis-Leistungs-Verhältnis lassen offensichtlich keine zuverlässige Differenzierung zu.

Diskussion und Resümee

Von den oben dargestellten potentiellen Gründen für niedrige ICC-Werte dürften hauptsächlich zwei eine Rolle spielen: Teilweise ist die Varianz zwischen den Gruppen zu gering, um sinnvolle Unterscheidungen treffen zu können. Problematischer scheint jedoch zu sein, dass es in den Programmbeschreibungen zu den entsprechenden Kriterien zu wenige Informationen gibt, um ein genaues Urteil abgeben zu können. Niedrige ICC-Werte weisen wohl weniger auf Schwachstellen des Beurteilungssystems oder der Gutachter hin, sondern eher auf (Darstellungs-) Schwächen der Präventionsprogramme. Die Konsequenz dieser Tatsache sollte allerdings nicht sein, auf entsprechende Kriterien zu verzichten, sondern vielmehr die Anbieter solche Programme aufzufordern, alle entscheidungsrelevanten Informationen möglichst umfassend darzustellen. Ein zusätzlicher Grund für geringe Übereinstimmungswerte kann auch darin liegen, dass für die 56 Einzelkriterien nur 31 zusammenfassende Beurteilungsskalen erhoben und berechnet wurden, d.h. es wurden meistens mehrere Kriterien zu einer globalen Beurteilung zusammengefasst. Diese

Schwachstelle wird allerdings im echten Beurteilungsverfahren ausgeglichen, weil dort zu dem Kriterium inhaltliche, qualitative Aussagen erwartet werden.

Gut konzipierte und dokumentierte Programme können jedoch anhand des Kriterienkatalogs offenbar ausreichend objektiv und zuverlässig beurteilt werden. Voraussetzung für reliable Beurteilungen sind die fachliche Expertise der Beurteiler und eine einheitliche Verankerung der Kriterien, um zu einer hohen Urteilskonkordanz zu gelangen, was durch eine intensive Vorbereitung und Schulung erreicht werden kann (vgl. Wirtz & Caspar, 2002). Die Anwender des Beurteilungskatalogs sollten mit den theoretischen Grundlagen von Gewaltentstehung und – prävention vertraut sein und methodische Kenntnisse bezüglich Qualitätssicherung und Evaluation haben. Es ist jedem Auftraggeber für entsprechende Programme zu empfehlen, bei seinen Entscheidungen psychologischen Sachverstand hinzuzuziehen.

Literatur

- Domsch, M. & Gerpott, T.J. (1986). Zum Problem der Reliabilität von Organisationsklimamessungen. *Zeitschrift für Arbeits- und Organisationspsychologie*, 30, 116-124.
- Giesler, M. (2003). *Kreativität und organisationales Klima: Entwicklung und Validierung eines Fragebogens zur Erfassung von Kreativitäts- und Innovationsklima in Betrieben*. Münster: Waxmann.
- Preiser, S. (2006). Qualitätskriterien für Programme zur Gewaltprävention und Gewaltverminderung. In I. Hertzstiel, S. Blaschke, I. Loisch & C. Hanckel (Hrsg.), *Vom Nürnberger Trichter zum Laptop? Schule zwischen kognitivem und sozial-emotionalem Lernen*. Kongressbericht der 16. Bundeskonferenz 2004. Berichte aus der Schulpsychologie (S. 487-494). Bonn: Deutscher Psychologen Verlag.
- Preiser, S. & Wagner (2003). Gewaltprävention und Gewaltverminderung: Qualitätskriterien für Präventions- und Interventionsprogramme. *Report Psychologie*, 28, 660-666.
- Preiser, S. & Wermuth, S. (2005). Qualitätskontrolle von Qualitätsbeurteilungen - Die Analyse von Expertenübereinstimmungen. In Berufsverband Deutscher Psychologinnen und Psychologen (Hrsg.), *Jung sein, alt werden. Congress-CD zum Deutschen Psychologentag 2005 / Kongress für Angewandte Psychologie, 10.-12. November 2005 in Potsdam*.
- Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

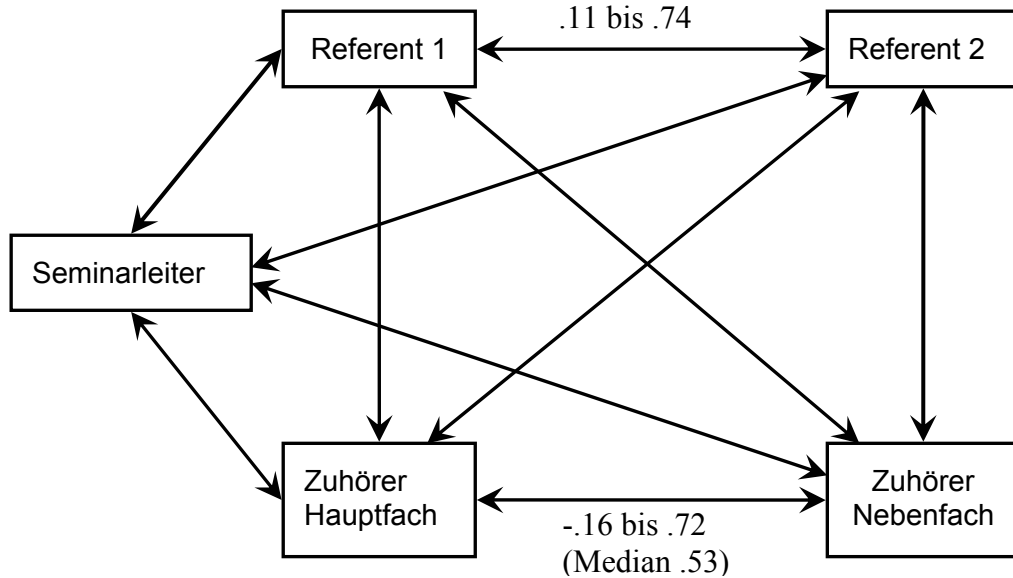


Abbildung 1: Ähnlichkeitsbeziehungen (Profilkorrelationen) für die Beurteilung von Präventionsprogrammen zwischen verschiedenen Beurteilern bzw. Beurteilergruppen (hier: Übereinstimmungen zwischen jeweils zwei Referenten sowie zwischen Hauptfach- und Nebenfachstudierenden) (aus: Preiser, 2006, S. 491)

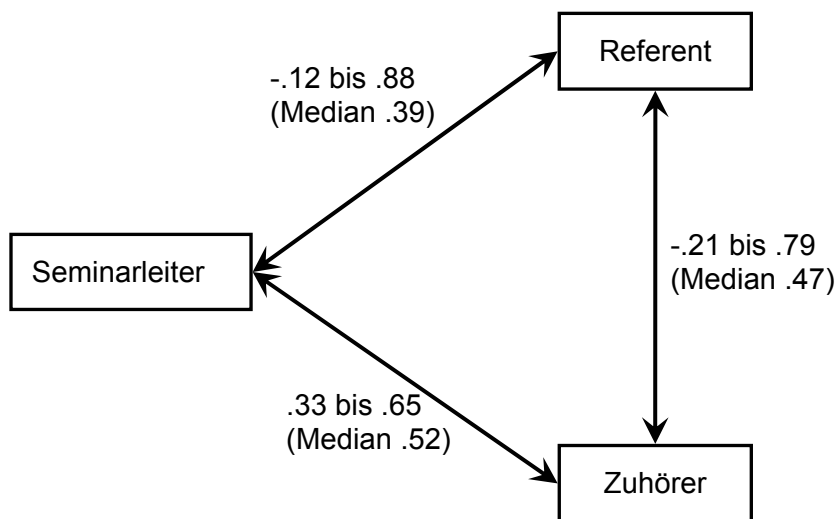


Abbildung 2: Ähnlichkeitsbeziehungen (Profilkorrelationen) für die Beurteilung von Präventionsprogrammen zwischen verschiedenen Beurteilern bzw. Beurteilergruppen (Die Korrelationskoeffizienten für mehrere Referenten sowie für Hauptfach- und Nebenfachzuhörer wurden zusammengefasst) (aus: Preiser, 2006, S. 492).

Anhang

Tabelle 1: Kennwerte der 7 Hauptdimensionen und der 31 Beurteilungsdimensionen des Qualitätskriterienkatalogs

Zielklärung: Benennung und Begründung konkreter und nachprüfbarer Ziele für die Teilnehmerinnen und Teilnehmer und für die beauftragende Institution ICC (1) = .18; ICC (2) = .40					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> Gibt es Aussagen über die Interventionsziele? Wie werden diese begründet? Sind die Ziele auf humanitäre und gesellschaftliche Wertvorstellungen bezogen? 	3,69	0,53	0,812	0,689	- 0,23
<ul style="list-style-type: none"> Gibt es Aussagen und Informationen über den Ausgangszustand (Ist-Zustand)? Gibt es klare Aussagen darüber, was konkret verändert werden soll (Soll-Zustand)? (Wissen, Einstellungen, Verhaltensweisen, grundlegende Kompetenzen, Schlüsselqualifikationen) 	3,51	0,77	2,544	0,006	0,61
<ul style="list-style-type: none"> Wird deutlich, anhand welcher nachprüfbarer Kriterien der Erfolg der Maßnahme überprüft werden kann? 	2,52	0,99	2,606	0,005	0,62
<ul style="list-style-type: none"> Werden realistische Effekte erwartet und quantifiziert bzw. präzisiert? 	2,82	0,95	0,844	0,654	- 0,18

Zielgruppe: Beschreibung der Zielgruppe mit Begründung des Bedarfs und der Erreichbarkeit ICC (1) = .14; ICC (2) = .35					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> Wird die Zielgruppe genau beschrieben? Wird begründet, warum bei dieser Zielgruppe ein bestimmter Bedarf besteht? 	3,58	0,71	1,475	0,147	0,32
<ul style="list-style-type: none"> Ist klar, wie die Zielgruppe erreicht werden kann? Ist die Teilnahme freiwillig oder verpflichtend? Werden mögliche Teilnahmehindernisse angesprochen? Ist geklärt, wie die Teilnehmergruppen zusammengesetzt werden sollen? 	3,57	0,58	1,203	0,303	0,17
<ul style="list-style-type: none"> Welche Annahmen oder Informationen gibt es über die Erwartungen und die Motivationslage der Zielgruppe? Was spricht aus deren Sicht für die Teilnahme? 	3,02	0,90	1,442	0,161	0,31

<ul style="list-style-type: none"> Wie werden erwartete Kompetenzen und Vorkenntnisse der Teilnehmenden berücksichtigt? 					
Theoretische Grundlagen: Explizite Benennung theoretischer Grundlagen für die geplanten Maßnahmen insgesamt und für die einzelnen Programmschritte; Bezugnahme auf empirisch gestützte Forschungs- und Anwendungsergebnisse <div style="text-align: right;">ICC (1) = .31; ICC (2) = .57</div>					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> Wird klar benannt, auf welche theoretischen Grundannahmen sich das Programm stützt? Sind diese Annahmen in sich schlüssig und kompatibel mit dem Forschungsstand? Werden die konkreten Interventionsziele aus diesen Grundlagen abgeleitet? 	3,08	0,88	1,302	0,236	0,23
<ul style="list-style-type: none"> Wird die Herkunft von Programmelementen in transparenter Weise dokumentiert? Werden die einzelnen Maßnahmen (Programmbausteine) aus den theoretischen Grundlagen abgeleitet? Wird auf empirisch gesicherte Erkenntnisse zur Wirksamkeit der Maßnahmen in Bezug auf die intendierten Ziele verwiesen? 	2,85	0,86	3,105	0,001	0,68

Maßnahmenbeschreibung: Beschreibung organisatorischer Rahmenbedingungen und konkreter Methoden und Medien; Teilnehmeraktivierung; Sicherstellung der Akzeptanz und der Teilnehmermotivation <div style="text-align: right;">ICC (1) = .25; ICC (2) = .51</div>					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> Gibt es Aussagen zu den Rahmenbedingungen (örtliche und räumliche Bedingungen, Gruppengröße, Zeitstruktur, Geräte- und Materialbedarf, Verpflegung, Unterbringung)? 	3,49	0,75	4,442	0,000	0,77
<ul style="list-style-type: none"> Werden unveränderliche Rahmenbedingungen berücksichtigt? Wird die Frage der Machbarkeit im jeweiligen Anwendungskontext beachtet? Wird die Kompatibilität mit den vorgegebenen Regeln und Strukturen des Anwendungsfeldes (z.B. Schule oder Strafvollzug) sicher gestellt? 	3,37	0,74	1,640	0,092	0,39
<ul style="list-style-type: none"> Welche Methoden kommen zum Einsatz? Wie werden deren erwartete Wirkungen begründet (verhaltens- und handlungsorientierte, themenzentrierte, 	3,44	0,82	3,225	0,001	0,69

kognitive, emotionale Methoden usw.) ?					
<ul style="list-style-type: none"> • Wie wird die Motivation der Teilnehmerinnen und Teilnehmer berücksichtigt und gefördert? Wie wird die Akzeptanz sicher gestellt? • Wie wird die aktive Beteiligung angeregt? Wie ist das Verhältnis von Forderungen an die Teilnehmer und Unterstützung? 	3,39	0,64	0,994	0,492	0,006
<ul style="list-style-type: none"> • Welche Medien kommen zum Einsatz? • Welche Materialien erhalten die Teilnehmerinnen und Teilnehmer zur Vor- und Nachbereitung? 	3,05	0,87	2,017	0,030	0,50
<ul style="list-style-type: none"> • Sind der Aufbau der Maßnahmen und die zeitliche Dauer und Struktur nachvollziehbar und begründet? • Werden die Maßnahmen ggf. mit gestaffelter Intensität – je nach Erfordernissen – angeboten? • Werden weiterführende Hilfs- oder Interventionsangebote spezifiziert? 	3,19	0,79	1,461	0,153	0,32
<ul style="list-style-type: none"> • Wird Flexibilität zwecks Fein-Anpassung an die Zielgruppe und die spezifische Problemlage eingeplant? Ist das Verfahren robust gegenüber individualisierten Modifikationen? 	3,37	0,74	1,247	0,272	0,20

Kompetenzen der Personen, die die Maßnahme durchführen:					
Nachweis der fachlichen und didaktischen Kompetenz der Trainerinnen und Trainer bzw. der Durchführenden					
ICC (1) = .39; ICC (2) = .66					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> • Welche fachlichen/wissenschaftlichen Qualifikationen haben die Trainer? 	3,42	0,79	1,800	0,058	0,44
<ul style="list-style-type: none"> • Welche didaktischen Erfahrungen oder Kompetenzen haben die Trainer? 	3,33	0,75	2,250	0,015	0,55
<ul style="list-style-type: none"> • Sind die Trainer mit dem System (z.B. Schule oder Strafvollzug) vertraut, in dem das Programm angewendet werden soll? 	3,47	0,67	1,978	0,034	0,49
<ul style="list-style-type: none"> • Wie werden die Anwender, Mediatoren oder Multiplikatoren des Programms ausgebildet, eingewiesen und supervidiert? • Wird die Maßnahme von Einzelpersonen oder einem Tandem/Team angeboten und durchgeführt? 	3,01	0,81	4,212	0,000	0,76

Evaluation und Qualitätssicherung: Integration einer systematischen Evaluation in die Programmentwicklung, - anwendung und -optimierung; Maßnahmen zur Qualitätssicherung ICC (1) = .09; ICC (2) = .25					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> • Welche Schritte zur Qualitätssicherung des Programms sind geplant? Wer ist dafür zuständig? 	2,65	0,81	0,767	0,738	- 0,3
<ul style="list-style-type: none"> • Ist Evaluation integraler Bestandteil der Maßnahme? Welche Evaluationsmethoden werden eingesetzt? • Ist eine Bewertung des Trainings und der Trainer vorgesehen? 	2,58	0,89	1,979	0,034	0,49
<ul style="list-style-type: none"> • Liefert die geplante Evaluation einen Soll-Ist-Vergleich? • Wie werden Veränderungen erfasst? • Wie werden die Ergebnisse der Evaluation bei der Weiterentwicklung der Konzeption berücksichtigt? 	2,44	0,93	1,044	0,442	0,04
<ul style="list-style-type: none"> • Werden vergleichbare Kontrollgruppen (die nicht an der Maßnahme teilnehmen) oder Wartekontrollgruppen (die erst zeitversetzt an der Maßnahme teilnehmen) berücksichtigt? 	1,94	1,27	4,802	0,000	0,79
<ul style="list-style-type: none"> • Ist eine zeitlich versetzte Abschlussevaluation (Nachbefragung) geplant? 	2,58	1,26	0,758	0,748	- 0,32
<ul style="list-style-type: none"> • Wie wird die Objektivität / Neutralität der Evaluation gesichert? 	2,25	0,97	1,325	0,222	0,25
<ul style="list-style-type: none"> • Welche konkreten Evaluationsergebnisse sind bereits dokumentiert? Sind sie zugänglich? • Welche Referenzen werden angegeben? 	2,56	1,13	2,334	0,012	0,57

Preis-Leistungs-Verhältnis (Effizienz):					
Angaben über Kosten, Nebenkosten und erwarteten Nutzen					
ICC (1) = -.14; ICC (2) = -.41					
Itemtext	M	SD	F	p	ICC(2)
<ul style="list-style-type: none"> • Welche Kosten und Nebenkosten entstehen? • Wo entstehen die Kosten? Wer ist Kostenträger? • Welche Kosten können durch das Programm eingespart werden? • Wie viele Teilnehmer werden durch die Maßnahme erreicht? • Was sind die Kosten pro Teilnehmer? 	2,98	0,74	1,170	0,330	0,15
<ul style="list-style-type: none"> • Welche Effekte werden erwartet, in welcher Höhe? • Wie ist die Breite der angestrebten Wirkung? Gibt es eine differentielle Wirksamkeit für bestimmte Personengruppen? 	2,47	0,81	0,567	0,916	- 0,76
<ul style="list-style-type: none"> • Werden Langzeiteffekte, Multiplikatoreffekte und positive Nebenwirkungen erwartet? 	2,84	0,84	0,735	0,771	- 0,36
<ul style="list-style-type: none"> • Werden Risiken oder potentielle negative Nebenwirkungen in der Planung berücksichtigt? 	1,97	0,94	0,885	0,609	- 0,13