Der NZK-Evidenzindex für das online Portal WESPE

Andreas Armborst

Präventionsansatz	Wirksamkeit	wissenschaftl. Belastbarkeit	Übertragbarkeit	Theoretische Fundierung	Anzahl Evaluationsstudien	Evidenzindex
Ambulante Behandlung	+				10	53
Stationäre Behandlung	+				18	64

Für die Beurteilung der Qualität wissenschaftlicher Studien gibt es unterschiedliche Bewertungsverfahren, wie beispielsweise die Maryland Scientific Methods Scale (SMS). Aufbauend auf diesen Instrumenten hat das NZK ein eigenes Bewertungsprotokoll für das online Portal WESPE entwickelt. Es liefert einen groben Anhaltspunkt für das wissenschaftliche Niveau von Evaluationen. Der Index hat Werte zwischen 0 bis 100, wobei 100 für eine wissenschaftlich exzellente Praxis- und 0 für ein gänzlich unwissenschaftliches Vorgehen bei der Evaluation stehen. Das Protokoll vereint einige der gängigen Bewertungskriterien aus der qualitativen, quantitativen und mixed methods Sozialforschung. Anhand des Bewertungsprotokolls wird der Evidenzindex gebildet.

Der Evidenzindex ist eine Bewertung in Bezug auf ein ganz konkretes Untersuchungsziel; nämlich der empirischen Evaluation von Präventionsmaßnahmen. Er ist ausdrücklich kein Maßstab zur Beurteilung der allgemeinen wissenschaftlichen Relevanz einer Studie. Auch eine Studie mit relativ niedrigem Evidenzindex kann durchaus einen wichtigen theoretischen und/oder empirischen Beitrag in Ihrer Disziplin leisten.

Der Evidenzindex dient dem NZK dazu Evaluationsstudien nach einem festgelegten Protokoll hinsichtlich ihrer Aussagekraft und empirischen Robustheit zu bewerten. Die Frage nach der Kausalität (Wirksamkeit) einer Präventionsmaßnahme ist dabei z.B. eines von insgesamt neun Bewertungsmerkmalen. Dadurch können auch Evaluationen, die nicht den rigorosen Anforderungen von experimentellen Studien (randomised controlled trials *RCT*s) entsprechen eine hohe Punktzahl erreichen.

Für den Nutzer des online Portals WESPE kann es sinnvoll sein die Bewertung einer Studie in den einzelnen Bereichen nachzuvollziehen. Dies erlaubt einen detaillierten Einblick über die Stärken und Schwächen der Evaluation. Zu einer Übersicht zu den einzelnen Studienprotokollen im Portal gelangt man, indem man auf die Zahl in der Spalte *Anzahl der Evaluationsstudien* klickt. Bewertungsmerkmale, die für Wirkungsevaluationen besonders relevant sind, sind in den beiden Spalten Wissenschaftliche Belastbarkeit (= interne Validität) und Übertragbarkeit (= externe Validität) separat ausgewiesen.

Bewertungsmerkmal	Bewertungsstufen
1. Ziele der Studie Die Formulierung der Untersuchungsziele wird mit einer von 3 Stufen beurteilt:	**Stufe 0:** Die Studie benennt kein klares oder kein überprüfbares Untersuchungsziel. **Stufe 1:** Die Studie benennt ein klares überprüfbares Untersuchungsziel (oder dieses Ziel ist offensichtlich), beschreibt aber nicht dessen methodische Operationalisierung. **Stufe 2:** Die Studie benennt ein klares überprüfbares Untersuchungsziel und beschreibt die Methode.
2. Eignung des methodischen Zugangs Die Eignung des methodischen Zugangs wird mit einer von 3 Stufen beurteilt:	**Stufe 0:** Ungeeignet. Der methodische Zugang ist offensichtlich ungeeignet für das Untersuchungsziel der Studie und die Studie unternimmt keinen überzeugenden Versuch den gewählten methodischen Zugang zu begründen. **Stufe 1:** Geeignet und unterlegen. Der methodische Zugang ist prinzipiell geeignet für das Untersuchungsziel der Studie, gleichzeitig wäre ein anderes Studiendesign geeigneter gewesen. **Stufe 2:** Geeignet und überlegen. Der methodische Zugang ist geeignet für das Untersuchungsziel der Studie. Es gibt keine (oder kaum) evaluationspraktische Hindernisse bei der Evaluation.
3.Theoretische Grundlagen Die Eignung der theoretischen Grundlagen der Evaluation wird mit einer von 3 Stufen beurteilt:	**Stufe 0:** Die Studie benennt keine theoretischen Annahmen über die Wirkungsweise der Präventionsmaßnahme (Black Box Evaluation). **Stufe 1:** Die Studie benennt theoretische Annahmen über die Wirkungsweise der Präventionsmaßnahme, stellt aber keinen (ausreichenden) Bezug zu dessen empirischen Überprüfung her. **Stufe 2:** Die Studie benennt theoretische Annahmen über die Wirkungsweise der Präventionsmaßnahme und stellt einen ausreichenden Bezug zu deren empirischen Überprüfung her.
4. Interne Validität Die **interne** Validität wird mit einer von 6 Stufen beurteilt.	**Stufe 0:** Ein Erhebungszeitpunkt, keine Vergleichsgruppe **Stufe 1:** ► Mind. zwei Erhebungszeitpunkte (Vor und nach der Intervention), keine Vergleichsgruppe. Oder: ► Ein Erhebungszeitpunk, mind. eine (nicht äquivalente) Vergleichsgruppe, bei der relevante Einflussfaktoren mit geeigneten multivariaten statistischen Methoden kontrolliert wurden. Z.B. cross-sectional studies mit Auswertung durch multivariate Verfahren (multiple Regressionsmodelle, Mehrebenenanalyse, SEM **Stufe 2:** Mind. zwei Erhebungszeitpunkte (Vor und nach der Intervention), mind. eine Vergleichsgruppe (fixed sample), ohne multivariate statistische Kontrolle **Stufe 3:** ► Mind. zwei Erhebungszeitpunkte (Vor und nach einer Intervention), mind. eine Vergleichsgruppe für die Äquivalenz per nicht-randomisierter Zuweisungstechnik hergestellt wurde (Matching, Parallelisierung) Oder: ► Mind. zwei Erhebungszeitpunkte (Vor und nach einer Intervention), mind. eine Vergleichsgruppe ohne Randomisierung, für die relevante Einflussfaktoren mit geeigneten multivariaten statistischen Methoden kontrolliert wurden. Z.B. fixed-sample Panel Design (Längsschnittstudien), Kohortenstudien, Ex Post Facto Control Group Design mit Auswertung durch multivariate Verfahren (Regressionsmodelle, Mehrebenenanalyse, SEM). Oder: ► Design wie in Stufe 4 und 5 mit Stichproben < 30 Personen pro Ver-

	suchsgruppe. **Stufe 4:** Mind. zwei Erhebungszeitpunkte (Vor und nach einer Intervention), mind. eine Vergleichsgruppe bei randomisierter Zuteilung der Teilnehmer mit leicht ergebnisverzerrenden Einschränkungen (z.B. differentieller Dropout). **Stufe 5:** Mind. 2 Erhebungszeitpunkte (Vor und nach einer gesetzten Intervention), mind. eine Vergleichsgruppe bei randomisierter Zuweisung ohne Einschränkungen.
5. Externe Validität Die **externe** Validität wird mit einer von 6 Stufen beurteilt.	**Stufe 0:** Die Stichprobe ist nicht repräsentativ: Die Ergebnisse sind nur für die Teilnehmer der Studie gültig **Stufe 1:** Die Stichprobe ist nicht repräsentativ, aber es gibt gut begründete theoretische Annahmen, dass die Ergebnisse der Studie auch auf eine Population oder auf einen Kontext mit ähnlichen Merkmalen wie in der Stichprobe übertragbar sind (z.B. Evaluationsergebnisse aus anderen Ländern, Evaluationsergebnisse einer „typischen" Schulklasse, einer typischen JVA etc. **Stufe 2:** ► Die Ergebnisse sind repräsentativ (durch Stichprobe oder Vollerhebung) für eine kleine Teilgruppe (bspw. alle Schüler zwischen 14 und 17 einer bestimmten Schule) innerhalb des gesamten Adressatenkreises der evaluierten Maßnahme (Schüler zwischen 14 und 17 bundesweit). Aufgrund der Spezifika der Stichprobe (eine bestimmte Schule in einer bestimmten Stadt innerhalb eines bestimmten Bundeslandes) lassen sich Ergebnisse aber nur sehr eingeschränkt auf andere Populationen innerhalb der gleichen Zielgruppe übertragen Oder: ► Stichprobe ist repräsentativ für eine größere Teilgruppe innerhalb des gesamten Adressatenkreises der evaluierten Maßnahme (bspw. alle Schüler ab 15 Jahre in NRW), aber es gab Stichprobenverzerrende Probleme beim Samplingverfahren (bspw. zu geringe Stichprobe (n< 30), zu geringe Teilnehmerquoten oder großer Stichproben-Fehler **Stufe 3:** Stichprobe ist repräsentativ für den überwiegenden (bezifferbaren) Teil des gesamten Adressatenkreises der evaluierten Maßnahme. Beispiel: die Stichprobe zielt eigentlich auf eine andere Grundgesamtheit (bspw. alle Schüler in NRW), deckt damit aber gleichzeitig den Adressatenkreis der Maßnahme (bspw. stadtteilbasiertes Präventionsprogramm) (überwiegend) mit ab. **Stufe 4:** Vollerhebung einer größeren Teilgruppe innerhalb des gesamten Adressatenkreises der evaluierten Maßnahme. Beispiel: Vollerhebung aller Schüler in NRW deckt gleichzeitig den Adressatenkreis der Maßnahme (bspw. stadtteilbasiertes Präventionsprogramm) (überwiegend) mit ab **Stufe 5:** Stichprobe (oder Vollerhebung) ist repräsentativ für den gesamten Adressatenkreis der Maßnahme auf dem gesamten Bundesgebiet
6. Messvalidität (Konstruktvalidität) Die Eignung der Messvalidität (Konstruktvalidität) der Evaluation wird mit einer von 6 Stufen beurteilt:	**Stufe 0:** ► Die Studie macht keine oder unzureichende Angaben darüber, wie die Effektgrößen operationalisiert sind. **Oder:** ► Die verwendeten Indikatoren oder Daten sind ungeeignet die Effektgrößen zu messen. **Stufe 1:** Augenscheinliche Validität - Die Studie verwendeten augenscheinlich aussagekräftige Indikatoren (und/oder kodifizierte Verfahren) zur Messung der Effektgröße, jedoch sind die Indikatoren weder theoretisch noch empirisch begründet. **Stufe 2:** Augenscheinliche Validität und theoretische Validität - Die Studie

	verwendeten augenscheinlich gute und theoretisch begründete Indikatoren (und/oder kodifizierte Verfahren) zur Messung der Effektgröße **Stufe 3:** Augenscheinliche Validität und Reliabilität - Die Studie verwendet augenscheinlich gute Indikatoren zur Messung der Effektgröße und kann deren Reliabilität empirisch nachweisen (bspw. Faktorenanalyse, Reliabilitätstest, Intercoder- Reliabilität etc.). **Stufe 4:** Die Studie verwendet ein theoretisch verankertes und empirisch bewährtes Messinstrument zur Messung der Effektgröße (die Reliabilität des Instruments wurde empirisch überprüft). **Stufe 5:** Multimethodal - Die Studie verwendet theoretisch verankerte und empirisch bewährte Messinstrumente und setzt mindestens zwei verschiedenen Methoden zur Messung der Effektgröße ein (ein Instrument mindestens Stufe 3 und ein weiteres Instrument mindestens Stufe 1).
7. Qualität der Datenauswertung Die Qualität der Datenauswertung und ihrer Dokumentation wird mit einer von 6 Stufen beurteilt:	**Stufe 0:** ► Die Studie beschreibt die Methoden der Datenerhebung und der Datenauswertung nicht. Der Leser kann nicht beurteilen, ob die angewendeten Verfahren für die Evaluation geeignet sind (Verletzung der Methodentransparenz). Die Autoren liefern auch auf Nachfrage keine Informationen zur Methodik der Evaluation. **Oder:** ► Die empirischen Daten und die qualitativen und/oder quantitativen Auswertungsverfahren passen offensichtlich nicht zueinander und es fehlt eine Begründung warum die Studie die erhobenen Daten ausgerechnet mit diesen Verfahren auswertet. **Stufe 1:** Die Auswertung ist lückenhaft oder unverständlich dokumentiert, erscheint nach Einschätzung der gegebenen Informationen aber plausibel. **Stufe 2:** Die Auswertung ist nachvollziehbar dokumentiert und augenscheinlich plausibel. Es werden aber wichtige Annahmen/Voraussetzungen zur Anwendung der Methode verletzt (z. B. zu geringer Stichprobenumfang, fehlerhafte Annahmen über die Verteilung, oder Skalenmaße. Bei Qualitative Auswertungsmethoden erfolgt keine Interpretation in Gruppen und/oder Inter-coder Analyse. Eine kritische Auseinandersetzung über die Einschränkungen (Verletzung von Annahmen und/oder fehlende Intercoderanalyse) erfolgt nicht, oder ist offenkundig angreifbar. **Stufe 3:** wie Stufe 2, aber es erfolgt eine kritische und offene Auseinandersetzung über die Verletzung von Annahmen sowie deren Auswirkung auf die Ergebnisse. Die Dargestellten Ergebnisse sind bei Berücksichtigung ihrer offen dargelegten Einschränkungen nicht weiter angreifbar. **Stufe 4:** Die Auswertung der empirischen Daten ist nachvollziehbar dokumentiert, erfolgt anhand der geeignetsten Auswertungsmethoden und ohne erkennbaren Fehler. **Stufe 5:** Wie Stufe 4 + multimodale Auswertung bzw. Triangulation.
8. Ergebnisinterpretation Die Qualität der Ergebnisinterpretation wird mit einer von 3 Stufen beurteilt:	**Stufe 0:** Unangemessene, verzerrte Interpretation der Ergebnisse. **Stufe 1:** Angemessene, reflektierte und sachliche Interpretation der Ergebnisse; unzureichende Diskussion möglicher Grenzen und Einschränkungen der Ergebnisse. **Stufe 2:** Angemessene, reflektierte und sachliche Interpretation der Ergebnisse; selbstkritische Reflexion möglicher Grenzen und Einschränkungen der Ergebnisse
9. Interressenkonflikte Das Bestehen möglicher Interessenkonflikte wird mit einer	**Stufe 0:** Studienautor ist Programmentwickler bzw. Mitarbeiter mit einem kommerziellen Interesse an der Vermarktung der Maßnahme. **Stufe 1:** Studienautor ist Programmentwickler bzw. Mitarbeiter, aber Pro-

von 3 Stufen beurteilt	gramm wurde (bisher) nicht kommerziell verbreitet, keine kommerzielle Vermarktung. **Stufe 2:** Unabhängige Evaluation durch externe Einrichtung/Person ohne erkennbaren Interessenkonflikt.
Abwertungsfaktor	Eine Studie kann in begründeteren Fällen abgewertet werden, wenn es gravierende Mängel und Einschränkungen gibt, die **nicht** in den 9 Bewertungskriterien berücksichtigt sind. Dazu wird der Evidenzindex mit einem Faktor zwischen 0.0 und 1.0 multipliziert.

1	Ziele der Studie	► Stufen: 0-2		
2	Eignung des methodischen Zugangs	► Stufen: 0-2		
3	Theoretische Grundlage	► Stufen: 0-2		
4	Interne Validität	► Stufen: 0-5	Summenindex	Evidenzindex
5	Externe Validität	► Stufen: 0-5	Wertebereich 0-30	Wertebereich 0-100
6	Messvalidität (Konstruktvalidität)	► Stufen: 0-5		
7	Qualität der Datenauswertung	► Stufen: 0-5		
8	Ergebnisinterpretation	► Stufen: 0-2		
9	Interessenkonflikte	► Stufen: 0-2		

Die Transformation vom Summenindex in den Evidenzindex erfolgt anhand der folgenden Formeln. Bewertungsmerkmale mit 6 Stufen sind dadurch höher gewichtet, als Bewertungsmerkmale mit drei Stufen.

$$I_{EVIDENZ} = \left(\frac{1}{N \cdot 30} \sum_{i=1}^{N} x_i \right) \cdot 100$$

N = Anzahl der berücksichtigen Evaluationsstudien

x_i = erzielte Gesamtpunktzahl der *n-ten* Studie geteilt durch maximal mögliche Gesamtpunktzahl (Wertbereich 0-30) (Percentage of maximum possible)

$$x_i = \frac{1}{30} \sum_{j=1}^{q} y_j$$

q = Anzahl der Qualitätskriterien (derzeit 9)

y_j = erzielte Bewertungsstufe einer Studie für das *j-te* Qualitätskriterium q

Quellen und Literatur

Beelmann, A. & Hercher, J. (2016). *Methodische Beurteilung von Evaluationsstudien im Bereich der Gewalt- & Kriminalitätsprävention: Beschreibung und Begründung eines Methodenprofils*. In Stiftung Deutsches Forum für Kriminalprävention (Hg.), Entwicklungsförderung & Gewaltprävention 2015/2016 - Aktuelle Beiträge aus Wissenschaft und Praxis (S. 97-116). Verfügbar unter: http://www.wegweiser-praevention.de/files/DFK/dfk-publikationen/2016_06_02_jahrbuch_wegweiser_2015.pdf

Groeger-Roth, F. & Hasenpusch, B. (2011): *Grüne Liste Prävention. Auswahl- und Bewertungskriterien für die CTC Programm Datenbank*. Landespräventionsrat Niedersachsen (Hrsg.). Verfügbar unter: http://www.gruene-liste-praevention.de/communities-that-care/Media/Grne_Liste_Bewertungskriterien.pdf

Farrington, D.P.; Gottfredson. D.C.; Sherman, L.W. & Wels, B.C. (2002). Maryland Scientific Methods Scale. In Sherman et al. (Hg.), *Evidence-Based Crime Prevention.* Routledge, New York, 13-21.