

Meta-Evaluation in der Beruflichen Bildung

Ein Fallbeispiel und die Einordnung dessen Befunde in eine übergreifende Analyse von Meta-Evaluationen

Dr. Stefan Silvestrini

Evaluation und Wirkungsforschung in der Berufsbildung

AG BFN-Forum, 30. November 2018, Wissenschaftszentrum Bonn



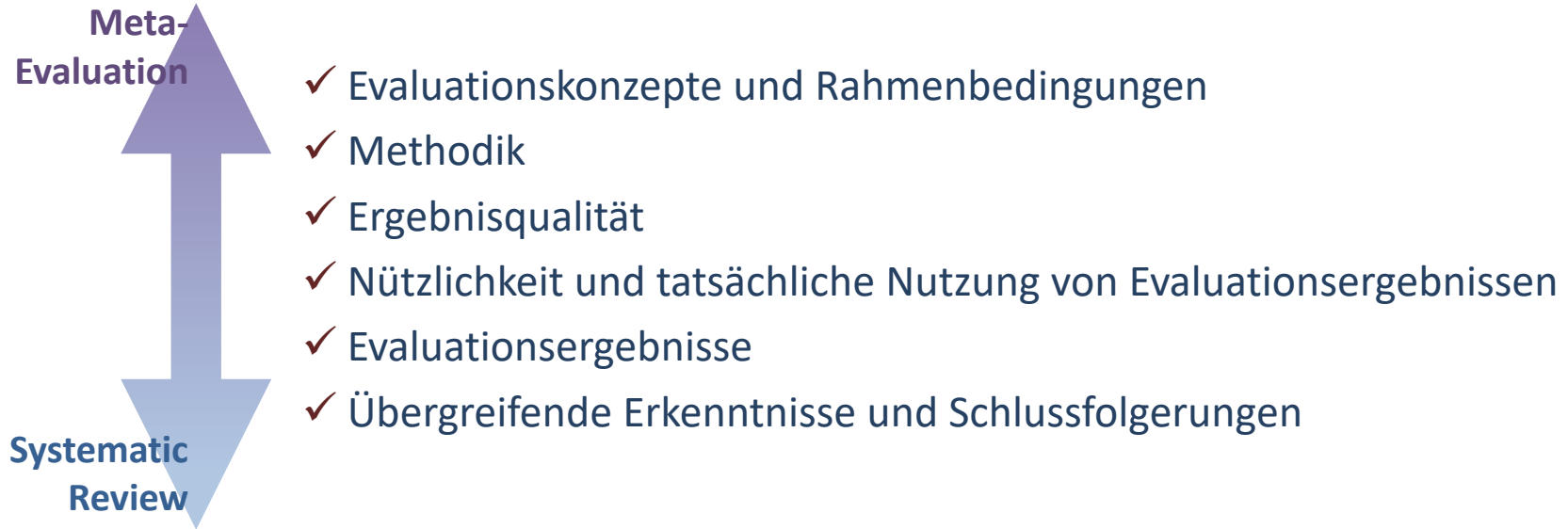
✓ Definitionen

- ✓ **Meta-evaluation:** “[...] the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation’s utility, feasibility, propriety, and accuracy and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide the evaluation and publicly report its strengths and weaknesses.” (cf. Stufflebeam 2001)
- ✓ **Systematic review:** “[...] the process of methodologically identifying, critically appraising, and selecting (high quality) evidence relevant to a specific research question and synthesising the results” (cf. Olsen and O-Reilly 2011)

✓ Meilensteine

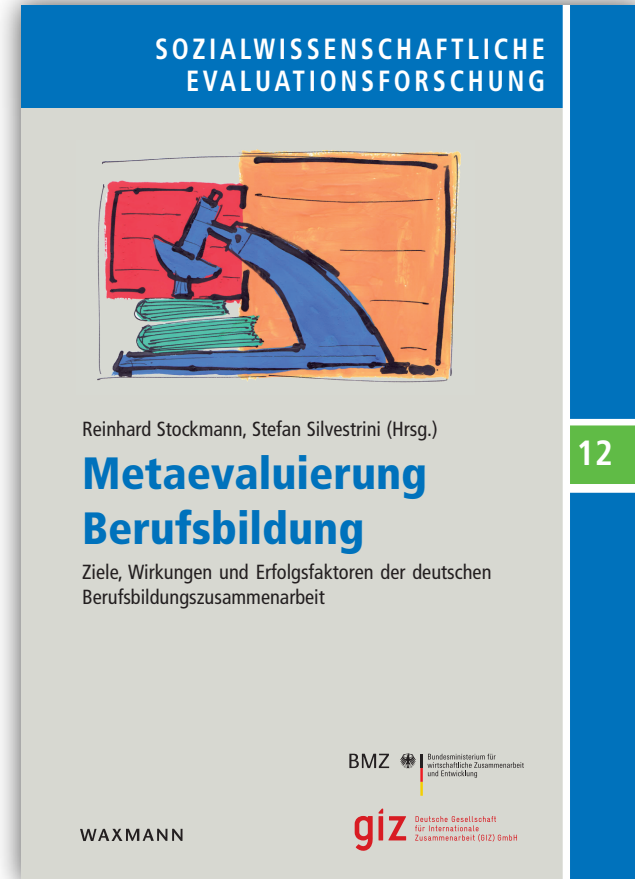
- ✓ 1981: Formulierung der JCSEE Program Evaluation Standards
- ✓ 1996: Thomas Widmer: Meta-Evaluation
- ✓ 2005: UNEG Normen und Standards für Evaluationen

- ✓ Offensichtlich steigender Bedarf an Meta-Studien
- ✓ Seit 2012 am CEval dreizehn Untersuchungen, die sich mit der Bewertung von Evaluationen und/oder der Synthese ihrer Ergebnisse befassen
- ✓ Untersuchungsdimensionen (u.a.):



- ✓ Wie können Evaluationen systematisch bewertet werden?
- ✓ Was sind typische Stärken und Schwächen von Evaluationsdesigns und ihrer Anwendung?
- ✓ Gibt es ‚typische‘ Evaluationsergebnisse, Schlussfolgerungen und Empfehlungen und wovon hängen diese ab?
- ✓ Gibt es übergreifende Faktoren, die den Erfolg bzw. die Wirksamkeit einer Intervention beeinflussen?
- ✓ Wo lohnt es sich weiter zu forschen?

Das Fallbeispiel



- ✓ Synthese
 - ✓ Zwölf unabhängige Evaluierungen aus 2010 und 2011
- ✓ Meta-Evaluierung
 - ✓ Zwölf o.g. Evaluierungsberichte
 - ✓ Zusätzlich 13 Berichte bestehend aus
 - ✓ Sieben Evaluierungsberichte von ex-InWEnt
 - ✓ Ein Evaluierungsbericht von ex-DED
 - ✓ Zwei Meta-Evaluierungen von InWEnt Vorhaben
 - ✓ Ein Evaluierungsbericht des BMZ

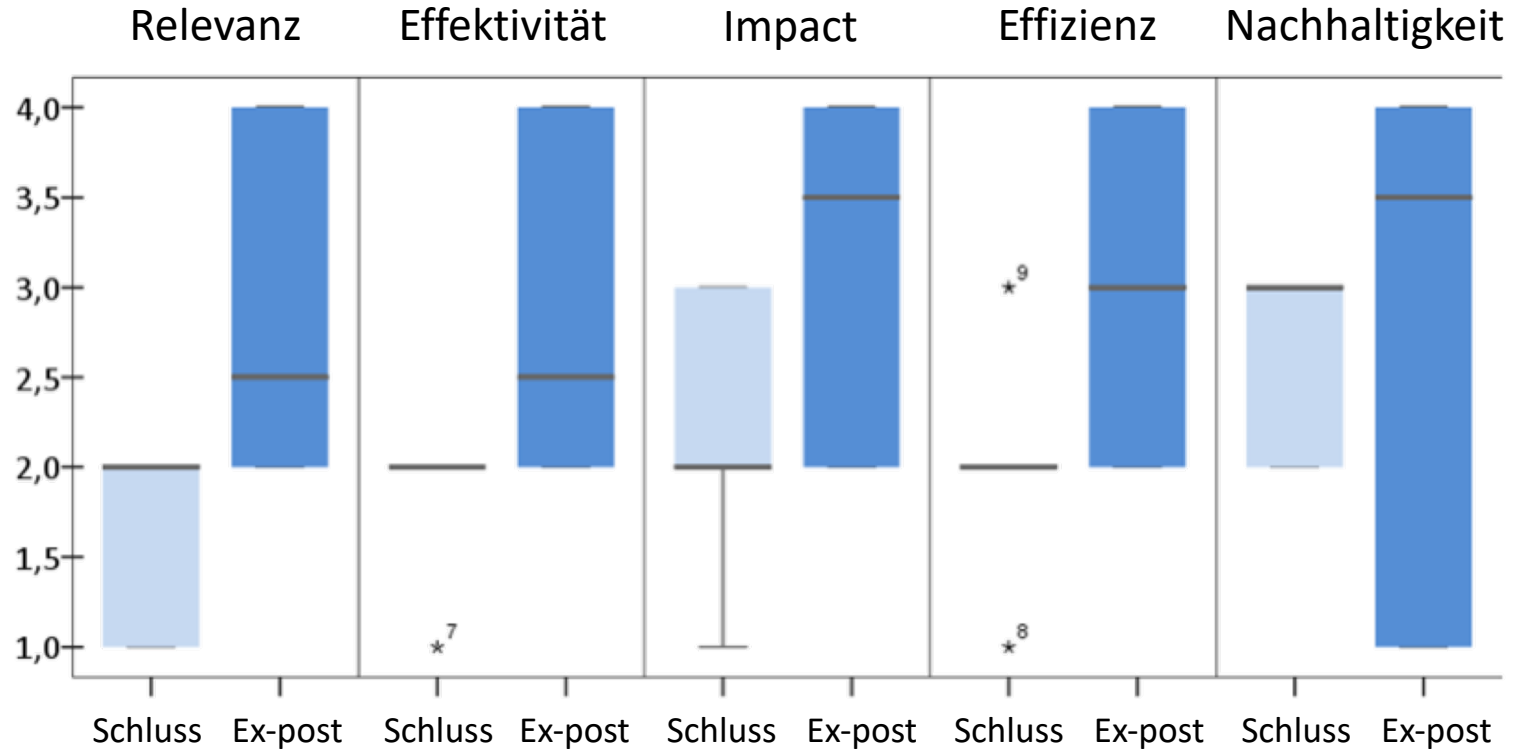
Nr. Land	Operative Kooperationspartner						Zielgruppen						Genderrifizierung	Fokussierung auf Armutsreduzierung
	Bildungs- einrichtungen	Übergeordnete Institutionen	Sonstige Behörden	Unternehmen	Verbände & Gewerkschftn.	Auszubildende	Arbeitssuchende	MA von Bildungseinr.	MA übergeord. Institutionen	Unternehmer	MA von Unternehmen			
1. Sierra Leone			✓	✓	✓	✓	✓			✓	✓	✓	✓	
2. Bulgarien	✓	✓			✓		✓	(✓)		✓	✓	✓	✓	
3. Indien	✓						✓	(✓)		✓	✓		✓	
4. Indonesien	✓	✓	✓			✓		✓	✓				✓	
5. Kasachstan	✓	✓			✓		✓	(✓)			✓	✓	✓	
6. Usbekistan	✓					✓		(✓)				✓		
7. Brasilien	✓	✓				✓		✓		✓	✓			
8. Burkina Faso			✓			✓			✓		✓	✓	✓	
9. China				✓		✓					✓	✓	✓	
10. Ghana	✓					✓	✓	(✓)		✓	✓	✓	✓	
11. Mazedonien	✓	✓					✓			✓			✓	
12. Philippinen	✓	✓		(✓)	✓	✓		✓	✓	(✓)		(✓)	✓	
Σ	9	6	3	2(3)	4	8	6	3(8)	3	5(7)	8	6(7)	10(12)	

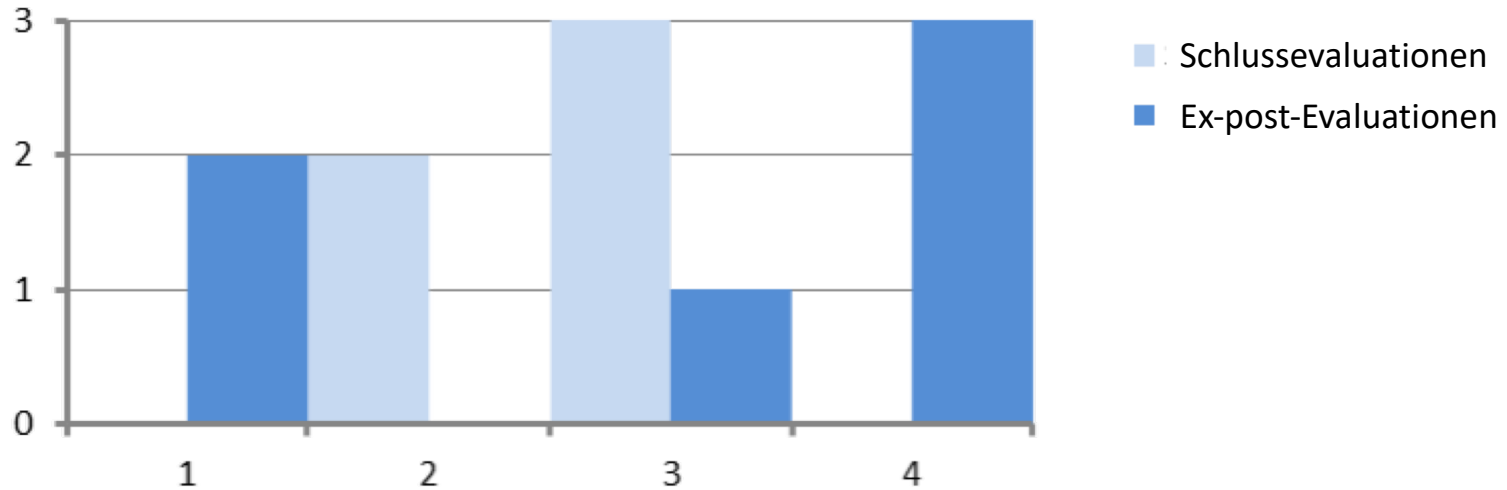
- ✓ Zusammenschau der Ergebnisse
- ✓ **Methodische Bewertung der Evaluationen**
- ✓ Historischer Vergleich
- ✓ Aufdeckung von Mustern und Trends
- ✓ Entwicklung vorhabenübergreifender Erkenntnisse
- ✓ Ableitung von Empfehlungen für neue Berufsbildungsvorhaben und integrierte Programmierung
- ✓ Verstärkte Nutzung der Evaluierungserkenntnisse

- ✓ Nutzung von Vergleichsgruppendaten in fünf Fällen
- ✓ Quasi-experimentelles Untersuchungsdesign in zwei Fällen
- ✓ Meist Kombination von qualitativen und quantitativen Methoden
- ✓ In allen Fällen erfolgt Daten- bzw. Quellentriangulation
- ✓ I.d.R. vollständige Beschreibung der Erhebungsinstrumente
- ✓ Immer „partizipative Vorgehensweise“
- ✗ Überwiegend Ex-post-facto Design
- ✗ Wenn Vergleichsgruppendesign dann keine Angaben zu Selektion der Vergleichsgruppe
- ✗ Kein systematischer Einbezug von Monitoringdaten
- ✗ Großzügige Verwendung des Begriffs „Multi-Methoden-Ansatz“
- ✗ Meist deskriptive Analysen, keine schließende Statistik
- ✗ Vereinzelte Verletzungen der Anonymität
- ✗ Keine Beschreibung der partizipativen Vorgehensweise
- ✗ Quantifizierung des Bewertungsergebnisses in Schulnotensystem subjektiv und daher nicht vergleichbar

Vergleichende Bewertung der DAC-Kriterien

	Typ	Dauer (Mon.)	Volumen (Mio.)	Rele- vanz	Effek- tivität	Im- pact	Effi- zienz	Nachhal- tigkeit	Gesamt- bewertung
1. Sierra Leone	Zwischen	54	3,00	1	2	2	3	3	2,2
2. Bulgarien	Schluss	70	2,00	2	1	1	2	2	1,6
3. Indien	Schluss	49	2,81	2	2	3	1	2	2,0
4. Indonesien	Schluss	59	9,39	1	2	2	3	3	2,2
5. Kasachstan	Schluss	41	3,25	2	2	3	2	3	2,4
6. Usbekistan	Schluss	85	3,08	1	2	2	2	3	2,0
7. Brasilien	Ex-post	42	1,16	2	2	2	3	1	2,0
8. Burkina Faso	Ex-post	77	5,38	4	4	4	4	4	4,0
9. China	Ex-post	119	6,39	2	3	2	2	1	2,0
10. Ghana	Ex-post	101	3,15	3	2	3	2	4	4,0 (2,8*)
11. Mazedonien	Ex-post	69	3,14	4	4	4	4	3	3,8
12. Philippinen	Ex-post	143	30,83	2	2	4	3	4	4,0 (3,0*)
Ø Schlussevaluationen		60,80	4,105	1,60	1,80	2,20	2,00	2,60	2,04
Ø Ex-post-Evaluationen		91,83	8,343	2,83	2,83	3,17	3,00	2,83	3,30
Ø insgesamt		75,75	6,848	2,17	2,33	2,67	2,58	2,75	2,68





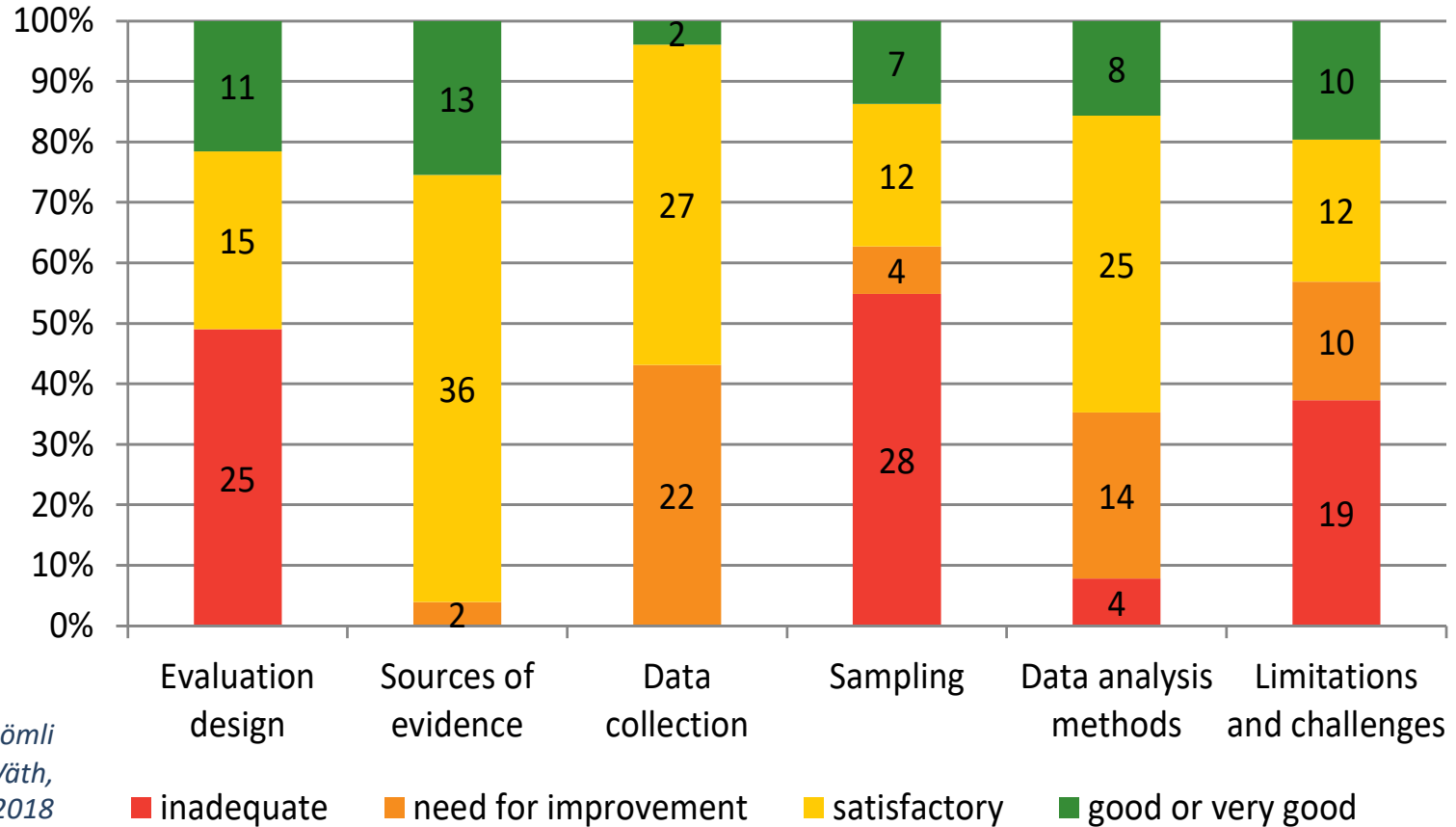
- ✓ Zur Nachhaltigkeitsbewertung ausschließlich Ex-post-Evaluierungen durchführen
- ✓ Berichtsgliederung hinsichtlich Verringerung von Redundanzen überarbeiten
- ✓ Vollständige und transparente Darstellung des Untersuchungsdesigns im Anhang der Evaluierungsberichte überprüfen, jedoch Angabe von personenbezogenen Daten vermeiden
- ✓ Verbesserung der Evaluierungsdesigns
- ✓ Einführung nachvollziehbarer und vergleichbarer Bewertungsregeln, einheitliche Berücksichtigung/Gewichtung des Einflusses der Rahmenbedingungen
- ✓ Reorganisation des Evaluationsprozesses
 - ✓ Intensiverer Dialog mit den Gutachtern
 - ✓ Zentrale ‚Qualitätskontrolle‘

Einordnung der Befunde – Meta²

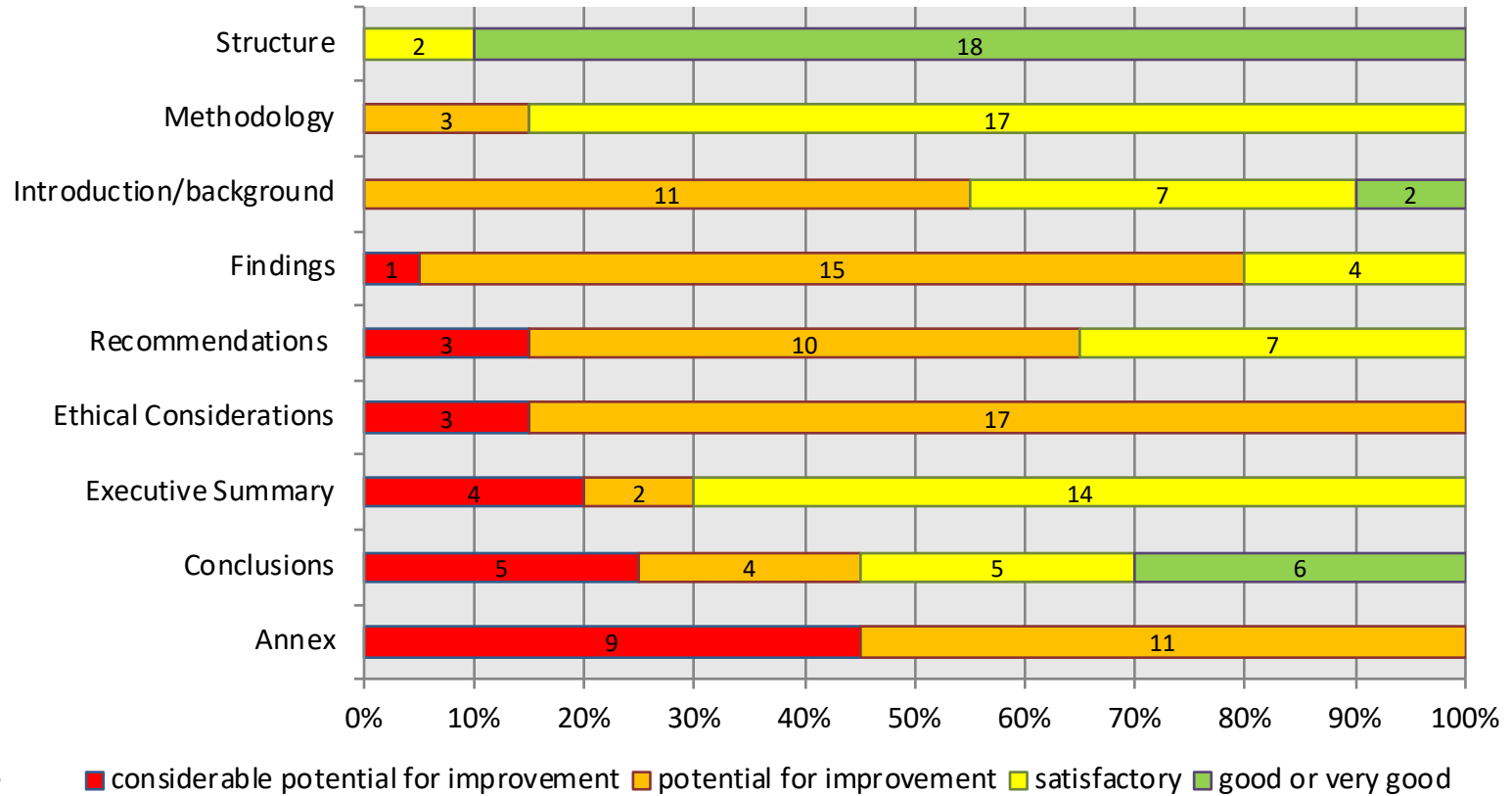
- ✓ Neun Meta-Studien auf Grundlage von rund 360 Projekt- und Programmevaluationen
- ✓ Unterschiedliche Sektoren der Internationalen und Entwicklungszusammenarbeit
- ✓ Vergleichbarer Untersuchungsansatz jedoch mit jeweils unterschiedlichem Fokus

Analysis dimension	Criterion	Scale	Rating guidance
Methodological assessment	Description of sample	1: insufficient 2: with considerable need for improvement 3: satisfactory, some need for improvement 4: good or very good	1: No description of sample 2: incomplete description, e.g. no ratio, size of basic population, composition 3: only description but no justification/rationale 4: Complete description including justification/rationale

- ✓ Offenbar Verbesserung der methodischen Qualität im Beobachtungszeitraum
- ✓ I.d.R. „good value for money“
- ✓ Meist umfassende und übersichtliche Ergebnisdarstellung
- ✓ Umfassende Dokumentation der praktischen Umsetzung
- ✓ Offenbar werden Evaluationsergebnisse als sehr nützlich angesehen
- ✗ Oftmals mangelnde Qualität der Leistungsbeschreibung
- ✗ Methodik und deren Umsetzung oftmals nur schwer nachzuvollziehen
- ✗ Vielen Evaluationen mangelt es an einer angemessenen Kontextanalyse
- ✗ Unzureichende Darstellung der Datenbasis, insbesondere bei Stichprobenziehung
- ✗ Methodische Limitationen und praktische Herausforderungen werden selten diskutiert
- ✗ Oftmals fehlende oder unzureichende Berücksichtigung konfundierender Faktoren bei Attributions- und Kontributionsanalysen
- ✗ Fehlende oder mangelhafte Effizienzbewertung
- ✗ Zuweilen Vermischung der Darstellung von Ergebnissen, Schlussfolgerungen und Empfehlungen



Liekefett/Mikkolainen/Römli
ng/ Silvestrini/Väth,
FMFA 2018



Bäthge/Väth,
CEval 2018

Relevanz

- ✓ Insgesamt das am besten bewertete DAC-Kriterium
- ✓ Hohe Übereinstimmung mit Strategien und Zielsetzungen der Partnerregierungen
- ✓ Bewertung korreliert hoch mit Bewertung der Wirksamkeit

Effektivität

- ✓ Auf Zielgruppenebene am besten bewertet
- ✓ Bewertung erfolgt meist auf Output-Ebene
- ✓ In etwa der Hälfte aller Evaluationen, in denen Gender-Effekte untersucht werden sollten, werden diese als nicht signifikant erachtet

Effizienz

- ✓ Meist ‚unter Vorbehalt‘ bewertet
- ✓ Positive Bewertungen liegen oftmals Synergieeffekte zugrunde
- ✓ Negative Bewertungen werden hingegen auf mangelnde Partnerkompetenzen und -beiträge, Personalfluktuaton und unzureichende Anpassungen an lokale Rahmenbedingungen zurückgeführt

Wirksamkeit

- ✓ Bewertung fokussiert vielfach auf intendierte Wirkungen
- ✓ Bisweilen werden auch bei kleinen, lokal begrenzten Interventionen Mutmaßungen über Beiträge auf nationaler Ebene angestellt
- ✓ Z.T. unzulässige Beschränkung auf soziale und wirtschaftliche Wirkungsdimensionen

Nachhaltigkeit

- ✓ Insgesamt das am schlechtesten bewertete DAC-Kriterium
- ✓ Überraschend oft auch bei Schlussevaluationen und z.T. sogar Zwischenevaluationen bewertet
- ✓ Oft werden mangelnde Partnerressourcen als Gefahr für Nachhaltigkeit genannt
- ✓ Disseminationseffekte bleiben oftmals außer Acht

Mehr M&E!

Beim nächsten Mal
Baseline-Daten erheben

Vergleichsgruppen-
Daten erheben

Volumen und/oder
Laufzeit erhöhen

Intervention weiter-
führen/ausweiten

Noch flexibler auf
Partnerbedarfe reagieren

Ownership der Partner
erhöhen

Kapazitäten der Partner
erhöhen

Partner weiterbilden

Partnerressourcen
mehr/besser nutzen

Mehr M&E!

Wirkungsindikatoren
überarbeiten

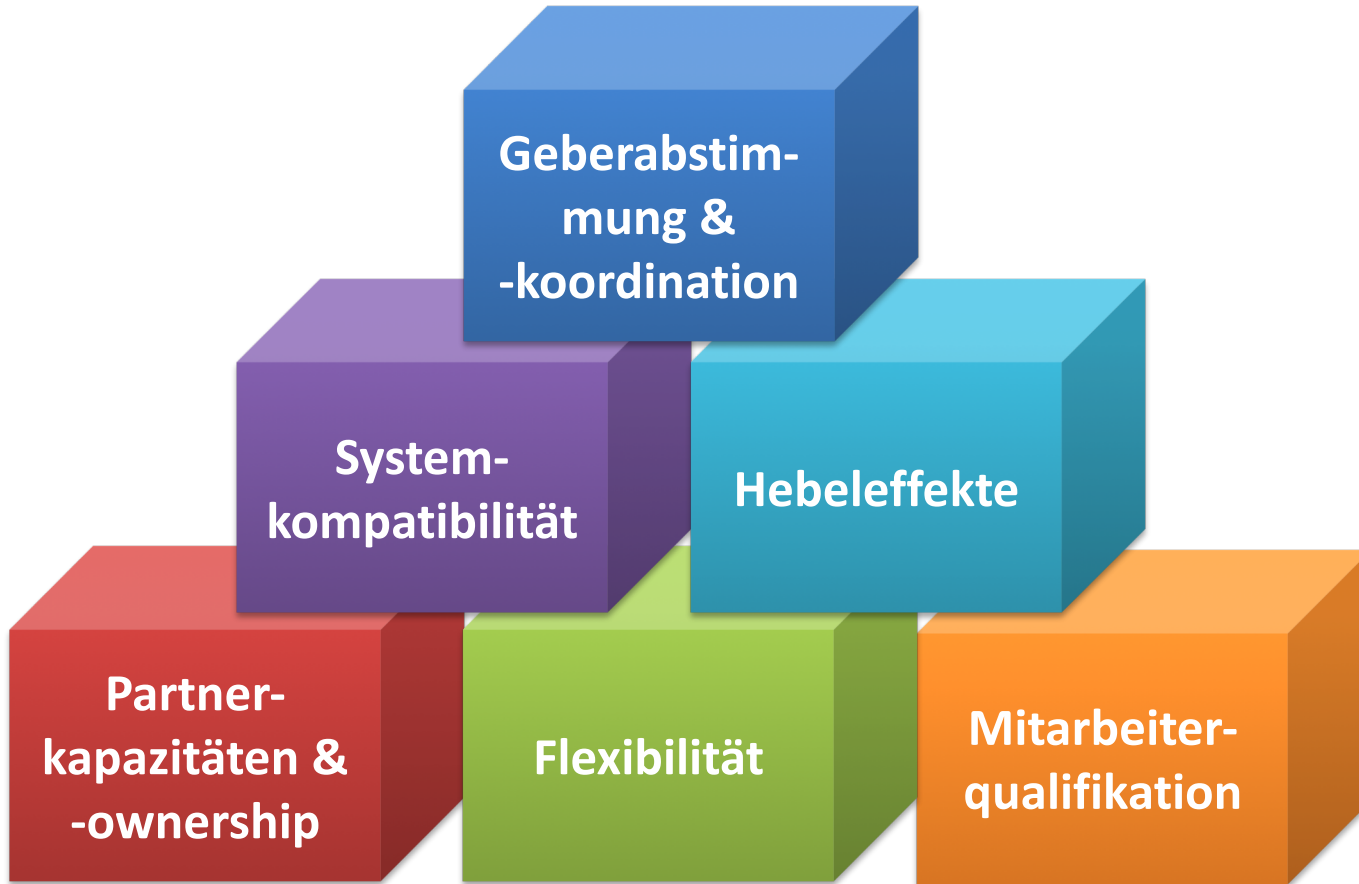
Wirkungsmodell
überarbeiten

Mehr M&E!

Geberabstimmung
verbessern

Mehr Zusammenarbeit
zwischen den Gebern

Mehr
Regierungsberatung



- ✓ In Anbetracht der Zahl sich wiederholender Befunde: Wieso wissen wir es nicht besser?
- ✓ Offensichtlich weisen Evaluationsergebnisse in unterschiedlichen Sektoren, Regionen viele Parallelen auf
- ✓ ... beeinflusst das Untersuchungsdesign die Evaluationsergebnisse
- ✓ ... gibt es weitere interventionsunabhängige Faktoren, die einen Einfluss auf die Ergebnisse haben

→ **Further research required!**

Vielen Dank für Ihre Aufmerksamkeit!

Dr. Stefan Silvestrini

T +49 (0)681 302 3679

E s.silvestrini@ceval.de

W www.ceval.de

Centrum für Evaluation



Center for Evaluation