

Vorläufer, Entstehung und Weiterentwicklung der Standards für Evaluation

1. Einleitung

Wenn man sie mit dem „typischen“ Entwicklungsverlauf eines neuen Tätigkeitsfelds bzw. einer entstehenden Profession vergleicht, haben Standards in der Evaluation eine überraschend lange Geschichte. Während ähnliche Regelwerke eher in späten Phasen von Professionalisierungsprozessen entwickelt werden, gehen unsere heutigen Evaluationsstandards auf Vorläufer zurück, die bereits in den 1970er Jahren erarbeitet wurden. Sie entstanden also zu einem Zeitpunkt, als die Tradition der gegenwärtigen Evaluationspraxis, deren Ursprung man in den großen gesellschaftlichen Reformen der 1960er Jahre sehen kann (Madaus & Stufflebeam, 2000), noch relativ jung war.

Vielleicht überrascht diese relativ frühe Auseinandersetzung mit Qualitätsansprüchen an das neu entstandene und wachsende Tätigkeitsfeld der Evaluation weniger, wenn man Folgendes bedenkt: Wir haben es hier mit einer einerseits methodisch sehr anspruchsvollen Aktivität zu tun, deren Akzeptanz andererseits in hohem Maße auf die Glaubwürdigkeit ihrer Ergebnisse angewiesen ist. Mit der steuernden Funktion nach Innen und dem Herstellen von Vertrauen nach Außen sind bereits zwei zentrale Funktionen von Standards angesprochen. Indem sie einen kollektiven Prozess des kumulativen Wissens- und Erfahrungsaufbaus seit Beginn der modernen Evaluationsgeschichte spiegeln und Gegenstand kontinuierlicher Aushandlungsprozesse sind, spielen sie darüber hinaus eine zentrale Rolle für die Identitätsbildung der Evaluationsgemeinschaft.

Ziel dieses einführenden Kapitels ist es, vor diesem Hintergrund einen übergreifenden Rahmen für die Auseinandersetzung mit den Standards für Evaluation zu schaffen, die in den weiteren Kapiteln des Bands erfolgt. Dazu beleuchten wir zunächst die oben angerissene Bedeutung von tätigkeitsbezogenen Standards und ähnlichen Regelwerken aus professionstheoretischer Sicht genauer (Abschnitt 2). Die Wurzeln der Standards für Evaluation in nordamerikanischen Vorarbeiten und deren Weiterentwicklung zu den Standards der schweizerischen SEVAL und der deutsch-österreichischen DeGEval zeichnen die Abschnitte 3 und 4 nach. Abschnitt 5 stellt den jüngeren Revisionsprozess der DeGEval-Standards dar, in Abschnitt 6 geben wir einen Ausblick vor dem Hintergrund internationaler Trends und Tendenzen in Bezug auf Evaluationsstandards.

2. Standards im Kontext der Professionalisierung der Evaluation

Fachspezifische Regelwerke wie Standards werden üblicherweise als zentrales und essenzielles Element der Professionalisierung einer Tätigkeit betrachtet. Historische Vor-

bilder in struktureller Hinsicht mag man bei religiös begründeten Codizes wie den zehn alttestamentarischen Geboten erkennen (Molander, 1987). Als ältestes Beispiel und Vorläufer aller auf eine konkrete Tätigkeit bezogenen, selbst auferlegten Prinzipien und Regeln kann der Eid des Hippokrates in der Medizin gelten (Hilton & Southgate, 2007), der allerdings eher dem Bereich der ethischen Verhaltenscodizes (ethical codes) zuzurechnen ist.

Obwohl im Begriffsfeld fachlicher Regelwerke (Beywl & Widmer, 2006) mit Elementen wie Standards, Verhaltenscodes (codes of conduct), ethische Leitlinien (ethical guidelines) oder Leitprinzipien (guiding principles) nur unscharfe semantische Grenzen und Differenzierungen existieren, gelten ‚Standards‘ im Evaluationskontext als bezüglich professioneller Praxis konkreter als ‚Verhaltenscodes‘ (codes of conduct) und spezifischer als Leitprinzipien (guiding principles) (Newman & Brown, 1996). Obwohl die Bezeichnung ‚Standards‘ keinen expliziten Bezug zu ethischen Aspekten professionellen Verhaltens nimmt, werden sie nicht als Gegenpol zu ethischen Leitlinien (ethical guidelines) verstanden, sondern schließen diese eher ein.

Will man die Rolle von Standards für die Professionalisierung der Evaluation diskutieren, ist zunächst eine weitere begriffliche Klärung vonnöten, da sowohl der Professionalitäts- als auch der Professionalisierungsbegriff mehrdeutig verwendet werden. Eine Aufgabe ‚professionell‘ zu erledigen kann einerseits meinen, dass sie nicht laienhaft, also ohne Erfahrung oder spezialisierte Expertise, durchgeführt wird. Andererseits kann v. a. in der alltagssprachlichen Verwendung des Begriffs gemeint sein, dass eine Aufgabe berufsmäßig oder als Teil einer beruflichen Tätigkeit durchgeführt wird, und nicht unentgeltlich als Amateur (Fox, 1992).

Entsprechend mehrdeutig ist die Forderung nach Professionalisierung in der Evaluation. Einerseits kann sie als Aufforderung zu einer möglichst fachgerechten Leistungserbringung verstanden werden, andererseits als Ruf nach Ausdifferenzierung eines neuen Berufsstands. Standards für Evaluation haben in beiden Zusammenhängen zentrale Rollen, die im Folgenden näher zu betrachten sind.

2.1 Standards zur Regulierung einer fachgerechten Leistungserbringung

Standards für Evaluation werden oft verkürzt als Handlungsanleitung für Evaluierende missverstanden. Auch wenn sie dies bis zu einem gewissen Grad leisten sollen, ist ihre Rolle aus professionstheoretischer Sicht jedoch deutlich vielfältiger. Nach Frankel (1989) lassen sich für Standards und ähnliche Regelwerke folgende Funktionen unterscheiden:

1. *Orientierung.* Als ‚Destillat‘ kollektiver Expertise können sie dem professionellen Handeln, besonderes in unklaren oder neuen Situationen, Orientierung und Unterstützung geben. Ähnlich einem Kompass geben sie eine Richtung vor, um besser fundierte Entscheidungen zu treffen, ohne das Ziel im konkreten Fall vorgeben zu wollen.
2. *Grundlage zur externen Bewertung.* Da sie selbstformulierte Ansprüche an die eigene Tätigkeit beinhalten, können Standards zur Bewertung der geleisteten

Arbeit und zum Zweck der Rechenschaftslegung dienen. Dies verweist auf die Rolle von Evaluationsstandards für die Meta-Evaluation, in der sie als Kriterien für die Bewertung der Güte oder Tauglichkeit von Evaluationsstudien dienen können (Scriven, 1969; Widmer, 1996).

3. *Professionelle Sozialisation.* Indem sie einen in der fachlichen Community allgemein akzeptierten und geteilten fachlichen Bezugspunkt darstellen, können Standards identitätsstiftend bzw. -stärkend und als Kristallisationspunkt für den Stolz auf die eigene Tätigkeit wirken.
4. *Quelle von Reputation und Vertrauen.* Professionalisierung bedingt eine zunehmende Autonomie des betroffenen Tätigkeitsfelds. Diese setzt auf der Gegenseite Vertrauen voraus, das mit Hilfe von Standards, auf die man sich verpflichtet und verpflichten lässt, gestärkt werden soll.
5. *Abschreckung unethischen Verhaltens.* Diese Funktion ist primär darauf angewiesen, dass (grobe) Verstöße gegen professionelle Standards gemeldet und sanktioniert werden. Im Fall von Evaluationsstandards erscheint dies bisher nicht realistisch, auch da es eine stärkere Operationalisierung voraussetzen würde, wann ein (grober) Regelverstoß vorliegt.
6. *Unterstützungssystem.* In Konfliktfällen können Standards vorbeugend als Argumentationshilfe gegenüber unangemessenen Forderungen oder Erwartungen dienen. Sie können Einzelnen den Rücken durch den Verweis auf das kollektive professionelle Selbstverständnis stärken (Balzer & Beywl, 2018).
7. *Konfliktbeilegung.* Schließlich können Standards rückwirkend in Konfliktfällen dazu dienen, Auseinandersetzungen bezüglich einer erbrachten Leistung zu schlichten und Meinungsverschiedenheiten einer fundierteren Klärung zuzuführen.

Die Standards für Evaluation sprechen in ihrem Selbstverständnis mehrere dieser Funktionen explizit an. Aber auch darüberhinausgehende Funktionen werden beansprucht. Primär ist dabei ihre Rolle für die Aus- und Weiterbildung in Evaluation zu nennen (vgl. den Beitrag von Schott, Niestroj und Friedrich in diesem Band). Zusätzlich sollen sie eine Referenz bilden, anhand dessen die Entwicklungsstufe der Evaluationspraxis in einem bestimmten (z. B. Länder-)Kontext beurteilt werden kann.

2.2 Standards als Schritt zur Herausbildung einer Profession

Mit der in den vergangenen Jahrzehnten gewachsenen Bedeutung der Evaluation ist ein neues Feld entstanden, in dem eine wachsende Anzahl von Personen teilweise oder sogar vorwiegend beruflich tätig ist. In der Vergangenheit lassen sich viele Beispiele finden, wie eine solche neue Tätigkeit zur Herausbildung eigener Berufsbilder und eigenständiger Professionen geführt hat.

Entsprechende Entwicklungsprozesse wurden in der Professionssoziologie anhand der klassischen Professionen wie Recht oder Medizin sowie neuerer Professionen wie Ingenieurwesen, Journalismus oder Psychotherapie untersucht (Mieg, 2010). Eine

idealtypische Abfolge von Entwicklungsschritten hat (Wilensky, 1964) auf Basis seiner Beobachtungen postuliert. Auch wenn deren Gültigkeit sich nicht pauschalisieren lässt (Abbott, 1991; Curnow & McGonigle, 2006), kann sie gut dazu dienen, die Rolle von Standards in diesem Kontext besser zu verstehen (Böttcher & Hense, 2015; Widmer, 2015).

1. Zunächst entsteht das neue Tätigkeitsfeld, und nur Einzelne befassen sich mit der Ausübung der neuen Tätigkeit.
2. Initiiert durch besonders engagierter ‚Pioniere‘ oder durch den Druck z.B. von Klienten oder einer interessierten Öffentlichkeit entstehen dann erste Aus- oder Fortbildungsprogramme.
3. Auf lokaler, regionaler, nationaler und/oder transnationaler Ebene schließen sich im neuen Tätigkeitsfeld Aktive in Fachverbänden zusammen. In dieser Phase kommt es Wilensky (1964) zufolge oft zu begleitenden Entwicklungen wie der Tendenz, Teilaufgaben der neuen Tätigkeit an weniger Qualifizierte zu delegieren. Es kann zu Konflikten zwischen Etablierten und Absolventinnen und Absolventen der neuen Ausbildungsgänge (s. Schritt 2) sowie zu Abgrenzungsbemühungen gegenüber verwandten Tätigkeitsfeldern kommen.
4. Bestrebungen der Abschottung gegenüber ‚nicht Professionellen‘ verstärken sich. Dies kann im Schützen der Berufsbezeichnung oder im Verbot einer nicht lizenzierten Ausübung der Tätigkeit münden, was eine entsprechende Lizenzierung oder Zertifizierung der qualifizierten Tätigkeitsausübung bedingt.
5. Schließlich werden Regeln etabliert, die professionelles Fehlverhalten sanktionieren, den innerprofessionellen Konkurrenzdruck mindern und dem Schutz von Klienten dienen sollen. Formal münden diese in professionellen Verhaltenskodexen oder Standards.

Dieses Phasenmodell muss nicht immer linear ablaufen, war aber bei der Mehrzahl der von (Wilensky, 1964) untersuchten Professionen zu beobachten. Es kann somit als Referenz dienen, um den aktuellen Professionalisierungsgrad der Evaluation zu diskutieren (Böttcher & Hense, 2015). An dieser Stelle interessiert aber vorwiegend die Rolle von Standards im Prozess der Professionsentwicklung. Teils unter dem Einfluss externer Anstöße (vgl. Abschnitt 3) wurden Standards relativ früh formuliert. Denn gemäß dem dargestellten Modell wäre deren Erarbeitung typischerweise erst nach einer formellen Schließung des Tätigkeitsfelds etwa durch die Etablierung zertifizierender Abschlüsse oder dem Schutz der Tätigkeitsbezeichnung zu erwarten.

In der Evaluation sind entsprechende Initiativen zur Schließung bisher nicht über fachinterne Diskussionen hinaus gekommen oder ergebnislos verlaufen (Altschuld, 1999). Einzig in Kanada konnte bisher ein – bisher allerdings rein freiwilliges – System zur Beglaubigung als Evaluator(in) erfolgreich installiert werden (Kuji-Shikatani, 2015). Dieser Sachverhalt mag vielleicht zum Teil erklären, dass auch die Standards für Evaluation in Bezug auf ihre Lenkungswirkung für die Praxis bisher schwach geblieben sind. Zwar hat man frühzeitig einen Meilenstein in Bezug auf Anforderungen an professionelle Evaluation erarbeitet. Trotz der differenzierten inhaltlichen Aussagen können sie gegenüber den von Wilensky (1964) beschriebenen Codes kaum expli-

zeit regulierend wirksam werden, da mit ihnen keinerlei Sanktionierungsmechanismen, etwa bei eklatanten Verstößen gegen die Standards guter Praxis, verbunden sind. Die Einhaltung der Standards bleibt somit zum jetzigen Zeitpunkt letztlich freiwillig. Sie sind zwar in einem befähigenden Sinne eine fachlich notwendige Grundlage für eine professionelle Evaluationsstätigkeit, aber sie dienen nicht der Schließung eines Berufsstandes.

3. Entwicklung der Standards des Joint Committee on Standards for Educational Evaluation

Die heutigen Evaluationsstandards der Gesellschaft für Evaluation beruhen auf zeitlich weit zurückreichenden Vorarbeiten:

Im Herbst 1975 nahm das Joint Committee on Standards for Educational Evaluation seine Tätigkeiten auf. Dieses Komitee vereinigte Vertretungen von US-amerikanischen Fachgesellschaften und Interessenorganisationen aus den Bereichen Psychologie, Bildung und Erziehung, darunter namentlich die American Educational Research Association (AERA), die American Federation of Teachers (AFT) sowie die American Psychological Association (APA). Führende Persönlichkeiten vertraten diese Organisationen, darunter etwa Donald T. Campbell (APA) und Egon G. Guba (AERA) (Joint Committee, 1981, S. viii). Das Joint Committee unter dem Vorsitz von Daniel L. Stufflebeam wurde unterstützt durch einen großen Kreis weiterer renommierter Fachpersonen, darunter Marvin Alkin, Ernest House, Michael Scriven, Robert Stake und Ralph W. Tyler. Deren Arbeiten prägen die Evaluation bis heute maßgeblich.

Wie aus der Bezeichnung hervorgeht, befasst sich dieses Joint Committee mit Standards für die Evaluation im Feld Bildung und Erziehung. Ein wichtiger Anstoß für die Aufnahme der Arbeiten ging von der APA aus. Das mit der Revision der APA-Standards for Educational and Psychological Tests and Manuals von 1966 befasste Komitee empfahl 1974 in seinem Schlussbericht, ein eigenständiges Standardwerk für die Evaluation von pädagogischen Programmen zu schaffen – einem wachsenden Einsatzfeld für psychologische Tests. Die Einrichtung eines zusätzlichen Komitees reagierte damit auf die quantitative Ausweitung und die methodologische Ausdifferenzierung der Evaluation, die sich seit den 1960er Jahren verstärkt mit der systematischen Beschreibung und Bewertung von Programmen, Projekten und Materialien in Bildung und Erziehung befasste (vgl. dazu ausführlich Beywl, 1988). Dies veranlasste die beteiligten Organisationen neben den Basisdokumenten namentlich zu psychologischen Test (American Psychological Association, 1966, 1974), an den Grundlagen zur Evaluation von Programmen, Projekten und Materialien zu arbeiten (Joint Committee, 1981).

Im Jahr 1981 veröffentlichte sodann das Joint Committee „Standards for evaluations of educational programs, projects, and materials“, eine 160-seitige Buchpublikation, die neben den dreißig Evaluationsstandards reichhaltige Zusatzinformationen und Hilfsmaterialien enthält (Joint Committee, 1981). Diese Standards sind gemäß den Merkmalen einer guten Evaluation, nämlich Nützlichkeit (utility), Durchführbarkeit (feasibility), Korrektheit (propriety) und Genauigkeit (accuracy) in vier Gruppen

organisiert. Diese überschaubare Vierergliederung ist sowohl für Evaluationsfachleute wie für Auftraggebende sowie weitere Stakeholder leicht vermittelbar und kann Bezugspunkt sein für den stets weiter zu führenden Diskurs über Güte, Tauglichkeit und allgemeiner, den Wert von Evaluationen. Sowohl bei der Formulierung der einzelnen Standards als auch der konkreteren Anforderungen wurde bereits bei dieser ersten Fassung auf methodologische Offenheit geachtet (Guba & Lincoln, 1989). Dies eröffnete breite Anwendungsmöglichkeiten bei Evaluationen mit unterschiedlichen Evaluationsverständnissen (Widmer, 1996).

Ging es bei dieser ersten Publikation um den Evaluationsgegenstand pädagogische Programme, so stellen nachfolgende komplementäre Buchpublikationen die Lernenden bzw. ihrer Lernleistungen in den Mittelpunkt (Joint Committee, 2003), und schließlich das in Bildungseinrichtungen tätige Personal (Joint Committee, 2009). Alle sind nach den vier Gruppen gegliedert.

Neben diesen auf Erziehung und Bildung ausgerichteten Arbeiten trieb die damalige Evaluation Research Society (ERS), eine der Vorläuferorganisationen der American Evaluation Association (AEA), die Entwicklung eines feldunspezifischen Sets von Standards voran. Diese Bemühungen resultierten 1982 in den „Evaluation Research Society standards for program evaluation“ (ERS Standards Committee, 1982). Diese etwas verspätet auch in deutscher Übersetzung (ERS Standards Committee, 1990) erschienenen Standards konnten sich jedoch nicht durchsetzen, wohl auch wegen einer methodologischen Engführung auf ein neopositivistisches beziehungsweise kritisch-rationales Evaluationsverständnis, das ein bestimmtes Ablaufmuster eines Evaluationsprozesses voraussetzte und damit emergente Evaluationsansätze ausgrenzte (Lincoln, 1985; Widmer, 1996).

Im Unterschied dazu erfreuten sich die Standards des Joint Committee in den 1980er Jahren großer Beachtung und breiter – wenn auch bei Weitem nicht flächendeckender – Anwendung (Widmer, 1996). Diese erfolgte auch über das Bildungsfeld hinaus.

Um aktuelle Entwicklungen in der Evaluation aufzugreifen, entschied sich das Joint Committee, im Jahr 1989 eine Überarbeitung der Standards von 1981 in Angriff zu nehmen (Joint Committee, 1994). In einem wiederum breit angelegten Entwicklungsprozess (Sanders, 2006) wurde die erste Auflage einem Revisionsprozess unterzogen. Dabei wurde die grundsätzliche Ausrichtung und namentlich die Gliederung in die vier Qualitätsdimensionen Nützlichkeit, Durchführbarkeit, Korrektheit und Genauigkeit beibehalten. Neben den Organisationen, die im Joint Committee bisher vertreten waren, kamen weitere hinzu. Es ergab sich mit der Canadian Society for the Study of Education eine geographische und mit der American Evaluation Association (AEA) eine thematische Erweiterung der Kommission, die nun von James R. Sanders geleitet wurde. Im Jahre 1994 konnte das Joint Committee sodann die neuen, nun als „The Program Evaluation Standards“ bezeichneten Evaluationsstandards veröffentlichen (Joint Committee, 1994). Das bereits 222 Seiten starke Buch umfasst wiederum dreißig Einzelstandards und weicht materiell nur geringfügig von der 1981er-Version ab. Verschiedene einzelne Standards wurden leicht angepasst, zwei wurden zusammengelegt, und ein neuer Standard zur Meta-Evaluation geschaffen (Widmer, 1996).

Diese zweite Auflage der Joint Committee Standards stieß auf ein sehr positives Echo und wurde in Nordamerika, aber auch in vielen anderen Regionen, und in Bildung und Erziehung, aber auch in vielen weiteren Evaluationsfeldern rege genutzt (Love & Russon, 2005; Yarbrough, Shulha & Caruthers, 2005; Russon, 2005; Widmer & Beywl, 2006). Sie dienten verschiedenen Evaluationsgesellschaften auch als Grundlage für eigene Standard-Entwicklungen (Widmer, 2005; Rouge, 2005; siehe unten, Kapitel 4).

Gut zehn Jahre nach der Publikation der zweiten Auflage leitete das Joint Committee (siehe <http://www.jcsee.org/>) eine neuerliche Revision der Program Evaluation Standards ein. Die daraus hervorgehende dritte Auflage erschien im Jahr 2011. Der neue Untertitel verweist auf eine Verschiebung hinsichtlich der intendierten Funktionen. Statt des bisherigen Untertitels „How to assess evaluations of educational programs“ (Joint Committee, 1994) wurde neu der Untertitel „A guide for evaluators and evaluation users“ (Joint Committee, 2011) gesetzt. Schon in den vorangehenden Auflagen gab es neben den prominenten orientierenden Teilen auch Leitlinien und Hinweise auf Fallstricke. Nun wird in der dritten Auflage die Handlungsorientierung nochmals deutlich gestärkt. Im Unterschied zu den bisherigen Versionen, welche das Schwergewicht auf die Darstellung von Merkmalen guter Evaluationen legen, behandeln diese neuesten Standards die Mittel ausführlicher, wie Evaluationsprozesse auszugestalten sind. Zudem hat sich das Joint Committee entschlossen, die bisherigen vier Qualitätsdimensionen Nützlichkeit, Durchführbarkeit, Korrektheit und Genauigkeit um die fünfte der Evaluationsverantwortlichkeit zu ergänzen (Joint Committee, 2011; Übersetzung der Kurzfassung siehe Beywl & Widmer, 2012).

Bis heute sind die Reaktionen der Evaluationsgemeinschaft in Nordamerika, aber auch in anderen Regionen auf die Program Evaluation Standards von 2011 eher verhalten. Mit der stärkeren Orientierung auf die einzusetzenden Mittel statt auf die wünschenswerten Qualitätsmerkmale einer Evaluation, büßen die Standards an methodologischer Offenheit ein. Sie sind daher womöglich für eine breite Evaluationsgemeinschaft, ungeachtet der verfolgten Evaluationsansätze oder -modelle, weniger als übergreifender Referenzrahmen geeignet. Das Dokument konkurriert mit der Formulierung von Handlungsempfehlungen (recommendations) für jeden Standard auch ohne zwingenden Grund mit Leitfäden oder Lehrbüchern für die Evaluationspraxis.

4. Von den JC-Standards zu den SEVAL- und DeGEval-Standards

Bis Ende der 1980er Jahre fanden die Standards des Joint Committee im deutschen Sprachraum wenig Aufmerksamkeit. Die Evaluation befand sich bis dahin in Deutschland, Österreich und der Schweiz noch in einem frühen Entwicklungsstadium. Es bestanden noch keine Fachgesellschaften und der fachliche Diskurs zur Evaluation befand sich noch in den Anfängen. Das lediglich als Band 1 erschienene Handbuch zur Evaluierungsforschung (Hellstern & Wollmann, 1984) griff die Arbeiten des Joint Committee nicht auf. Im Rahmen der Recherchen zu seiner Doktorarbeit stieß Wolfgang Beywl auf die Evaluationsstandards des Joint Committee von 1981 und nutzte

diese zur qualitativen Einschätzung des von ihm adaptierten und erprobten responsiven Evaluationsmodells (1988). Im Rahmen seiner Lehre und in der wissenschaftlichen Weiterbildung nutzte er die Joint Committee Standards als zentrales Referenzdokument und veröffentlichte Übersetzungen der Kurzfassung sowie der von der AEA beschlossenen „Guiding Principles for Evaluators“ (Newman, Scheirer, Shadish & Wye, 1995) zusammen mit einer Einordnung der beiden Dokumente. Seine Schlussfolgerung in Bezug auf die Evaluation im deutschsprachigen Raum: „Gerade weil der Professionalisierungsprozeß noch in den Kinderschuhen steckt, sollten die vorhandenen Orientierungshilfen aus dem angelsächsischen Sprachraum pragmatisch genutzt werden“ (Beywl, 1996, S. 92).

Parallel begann sich Thomas Widmer im Rahmen seiner Dissertation zur Meta-Evaluation (1996) mit der Frage der Evaluationsqualität auseinanderzusetzen. In der Folge erarbeitete er eine Übersetzung (und Erweiterung) der Joint Committee Standards von 1981 und setzte diese zur Qualitätsbewertung von zehn Evaluationen aus der Schweiz ein (Widmer, 1996). Eine ergänzende Studie nutzte diese Evaluationskriterien zur Erfassung der Qualität von kleinen Kurzevaluationen im Vergleich zu umfangreicheren Studien (Widmer, Rothmayr & Serdült, 1996).

Aufgrund dieser Befassung mit den Evaluationsstandards des Joint Committee und der Beobachtung, dass sich die Evaluation ab Mitte der 1990er auch im deutschsprachigen Raum allmählich ausbreitete, entschieden sich Beywl und Widmer, eine deutschsprachige Übersetzung der zweiten Auflage der Program Evaluation Standards (Joint Committee, 1994) zu erarbeiten. Sie wurde – zusammen mit begleitenden Analysen und Texten – als Buchpublikation veröffentlicht. Damit sollte für die deutschsprachige Evaluationspraxis der Zugang zu den Evaluationsstandards erleichtert werden. Befördert durch das stark steigende Interesse an Evaluationsfragen stieß dieses Handbuch der Evaluationsstandards (Sanders, 2006) auf großes Interesse, sodass nach der Erstveröffentlichung im Jahre 1999 im Jahr 2000 eine zweite und sodann im Jahr 2006 eine dritte Auflage erschien.

In der Schweiz entwickelte sich das Evaluationsfeld in den 1990er Jahren sehr dynamisch (Widmer, 2017). Gestützt auf zwei Forschungsprogramme, in deren Rahmen auch die Dissertation von Widmer entstanden ist, gründeten evaluationsinteressierte Personen aus Wissenschaft, Verwaltung und privaten Büros im Jahr 1996 die Schweizerische Evaluationsgesellschaft (SEVAL). Bald nach ihrer Gründung setzte der SEVAL-Vorstand auf Veranlassung des Gründungspräsidenten, Werner Bussmann, eine Arbeitsgruppe Evaluationsstandards ein und betraute Widmer mit deren Leitung. Gestützt auf seine Erfahrungen in der Anwendung der Joint Committee Standards und parallel zur laufenden Buchübersetzung konnte die Arbeitsgruppe dem SEVAL-Vorstand im Jahr 2000 eine bereinigte Fassung von Evaluationsstandards der Schweizerischen Evaluationsgesellschaft, kurz SEVAL-Standards, präsentieren. Diese wurde im Frühjahr 2001 durch die Mitgliederversammlung der SEVAL als Dokument verabschiedet (Widmer, Landert & Bachmann, 2000).

Diese SEVAL-Standards von 2000 beruhen weitgehend auf den Joint Committee Standards aus dem Jahr 1994, bei allerdings lediglich 27 Einzelstandards. Sie folgen in der Struktur den vier Qualitätsdimensionen Nützlichkeit, Durchführbarkeit, Kor-

rektheit und Genauigkeit. Methodologische Offenheit und die breite Anwendbarkeit waren bei der Entwicklung zentrale Gesichtspunkte. Die SEVAL-Standards etablierten sich in den Folgejahren als bedeutende Leitlinie für die schweizerische Evaluationsgemeinschaft, in der sie eine ausgesprochen hohe Bekanntheit und sehr häufige Verwendung fanden (Balzer, 2013; Widmer, 2011).

Trotz dieser positiven Entwicklungen entschied sich die Arbeitsgruppe Evaluationsstandards der SEVAL im Rahmen der Revisionsarbeiten in den Jahren 2013 bis 2016, vom bisherigen Ansatz Abstand zu nehmen (zum Revisionsverfahren der SEVAL vgl. den Beitrag von Rüefli und Zweidler in diesem Band). Anstelle der Strukturierung nach den vier Qualitätsdimensionen erfolgt die Gliederung der Standards neu nach einer Prozess-Ablaufstruktur, was einer methodologischen Engführung Vorschub leistet. Die SEVAL-Standards 2016 sind in drei Gruppen gegliedert: Allgemeine Grundprinzipien, Planung und Durchführung sowie Bewertung und Ergebnisvermittlung (Schweizerische Evaluationsgesellschaft [SEVAL], 2016). Die Revision reagierte damit auf namentlich von weniger Evaluationserfahrenen vorgebrachte Wünsche, das Dokument als Anleitung zur Evaluationsplanung zu gestalten. Seine Funktion als Orientierungs- und Reflexionsrahmen tritt damit zurück.

Für die Evaluation in Deutschland und Österreich begannen die Arbeiten an einem Standardset kurz nach der Gründung der DeGEval im Jahr 1997 unter explizitem Bezug auf die „Vorbilder“ Joint-Committee- und SEVAL-Standards (Beywl, 1999; Gesellschaft für Evaluation [DeGEval], 2002). Eine achtköpfige Kommission, bestehend aus sechs Personen von der Auftragnehmer- und zwei von der Auftraggeberseite erarbeitete mit Unterstützung von 13 Kommentatorinnen und Kommentatoren einen Entwurf. Dieser wurde auf Empfehlung des Vorstandes am 4. Oktober 2002 von der Mitgliederversammlung der DeGEval beschlossen. Von den SEVAL-Standards wurde die Ausrichtung auf methodologische Offenheit und Anwendbarkeit in breiten Evaluationsfeldern übernommen. Die Standardgruppe propriety wurde statt mit „Korrektheit“ mit „Fairness“ übersetzt (zu weiteren Unterschieden vgl. den Beitrag von Widmer 2006 sowie die im selben Band abgedruckte Transformationstabelle auf den Seiten 345–346). Die DeGEval-Standards fanden in den Folgejahren starke Verbreitung, besonders in Lehrbüchern und anderen Buchpublikationen zur Evaluation. Die DeGEval erarbeitete ihre weiteren „Empfehlungen“¹ für die Selbstevaluation, für die Aus- und Weiterbildung in Evaluation sowie für Auftraggebende von Evaluation unter enger Bezugnahme auf die Standards. Der DeGEval-Vorstand nimmt in seinen bislang zehn „Positionspapieren“² vielfach explizit Bezug auf die Evaluationsstandards. Eine Google-Suche nach „Nützlichkeit Durchführbarkeit Genauigkeit Korrektheit/Fairness Evaluation“ (4. Juli 2019) ergibt über 40.000 Einträge. Diese belegen, dass sich das Grundkonzept der Standard-„Familie“ Joint Committee/SEVAL/DeGEval weit ausgebreitet hat, sowohl in den wissenschaftlichen wie den politisch-administrativen Raum, als auch bei Trägern und Finanziers von Programmen, wie z. B. Stiftungen. Ausschreibungen in Deutschland oder Österreich nehmen häufig, dabei seltener als in der Schweiz, Bezug auf die Standards. Hinweise zur konkreten Nutzung der Standards in

1 www.degeval.org/publikationen/

2 www.degeval.org/publikationen/positionspapiere/

der Evaluationspraxis ergab auch die Onlinebefragung zum ersten Revisionsanlauf im Jahr 2004.

5. Revision der DeGEval-Standards

Standards sind selbstverständlich selbst Gegenstand von Bewertung: Werden sie tatsächlich als Referenz für gute Evaluation eingesetzt? Erweisen sie sich als nützlich, wenn ja, für wen und für welchen Zweck? Sind sie verständlich? Sind sie untereinander möglichst frei von Widersprüchen? Die Standards müssen sich als anpassungsfähig erweisen. Die obige Skizze ihrer Entwicklung bezeugt das. Auch die Standards der DeGEval stellen sich dieser Herausforderung.

5.1 Erster Revisionsanlauf 2004

Eine Revision der DeGEval-Standards wurde in der Einleitung der 1. Fassung bereits für 2004 in Aussicht gestellt. Neben ausführlicheren Erläuterungen wurden, ähnlich wie im Vorbild der Joint Committee Standards, auch Hinweise für die praktische Anwendung sowie Fehlanwendungen und auch Fallbeispiele in Aussicht gestellt.

Zur Orientierung und Vorbereitung der Revision wurde im Frühjahr 2004 eine Onlinebefragung durchgeführt. (vgl. ausführlich Beywl & Kotitschke, 2004a, 2004b). Daran haben sich ca. 240 Personen beteiligt, ca. die Hälfte der damaligen DeGEval-Mitglieder. Zentrale Fragestellungen waren die nach der Dringlichkeit einer Revision und der gewünschten Schwerpunktsetzung. Auffällig war, dass lediglich 20% der Antwortenden angeben, 75% oder einen größeren Anteil ihrer Tätigkeit auf Evaluation zu verwenden.

Ergebnisse zu den geschlossenen Fragen

Vertrautheit mit den und Wertschätzung der Evaluationsstandards waren beide ausgeprägt: 75% arbeiteten mit der Druck- oder Onlinefassung und ebenso viele hatten die Standards ein- oder mehrfach weiterempfohlen. Die Frage nach der gewünschten Ausrichtung der Revision zeichnete ein heterogenes Bild.

Dies wird in der Verteilung auf die vorgegebenen Antwortoptionen deutlich:

- (1) Eine Revision sollte sich ausschließlich auf den DeGEval-Standardtext (die 25 Einzelstandards) beziehen.
- (2) Eine Revision sollte sich ausschließlich auf die Erläuterungen beziehen.
- (3) Eine Revision sollte sich auf den gesamten Inhalt der Broschüre beziehen.
- (4) Eine Revision sollte in Inhalt und Umfang deutlich über die Broschüre „Standards für Evaluationen“ hinausgehen (wie in der Einleitung zur 1. Fassung skizziert).
- (5) Eine Revision der Broschüre halte ich z. Z. für nicht erforderlich.

Die beiden Positionen „Neues Produkt“ und „keine Revision“ markieren die Extreme. Ein Viertel sprach sich für eine mittelgroße Überarbeitung aus („Broschüre ge-

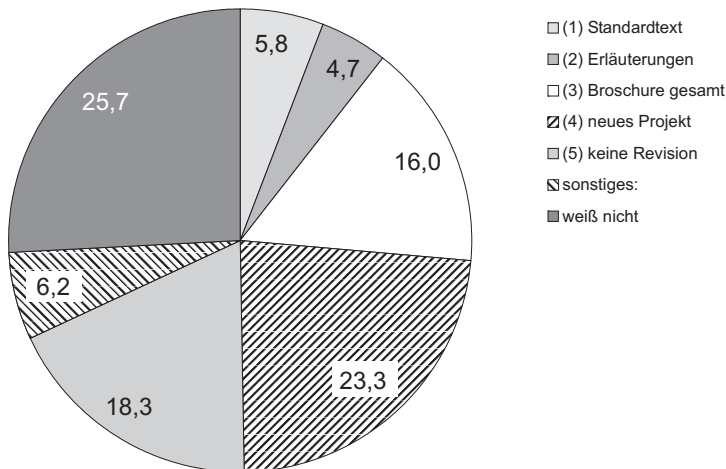


Abbildung 1: Revisionsalternativen (Quelle: Beywl & Kotitschke, 2004a, S. 13)

samt“ + „Erläuterungen“ + „Standardtext“), ein Viertel war unentschieden. Von den 144, die sich auf irgendeine Form der Revision festlegten, plädierte die größte Gruppe (ca. 47%) für den späteren Revisionsabschluss im Jahr 2008.

Ergebnisse zu den offenen Fragen

Diese fielen recht umfangreich und differenziert aus: Über 80% der Befragten (82,5%) antworteten auf mindestens eine der offenen Fragen. 508 Beiträge bringen es auf 106.874 Zeichen (ohne Leerzeichen), also ca. 40 A4-Seiten Text. Die Beiträge wurden mittels Text-Sortier-Technik (Beywl, Bestvater & Friedrich, 2011) in über 1.000 Sinn-einheiten zerlegt und zusammenfassend ausgewertet.

Etwa die Hälfte der Sinneinheiten bezieht sich auf verschiedene Nutzungen der DeGEval-Standards. In der Vertikalen sind diese danach angeordnet, inwiefern die Nutzung eher instrumentell, also für konkrete praktische Anwendungen erfolgt, oder eher konzeptionell, mithin für Reflexion, Theoriebildung oder allgemeinere Orientierung (Beywl, Künzli, Messmer & Streit, 2015). Im Mittelfeld (nach rechts) ist die Legitimierungsfunktion gegenüber Politik/Gesellschaft, bzw. (nach links) Bezug auf die Professionalisierung von Evaluation, auch in anderen Berufen/Disziplinen angesprochen.

Weitere ca. 300 Sinneinheiten liefern Begründungen für die jeweils gewählte Revisionsalternative (s. Abb. 1). Ungefähr 85 inhaltliche Verbesserung- und Erweiterungsvorschläge werden gemacht, z. B. bzgl. Fallbeispielen aus *verschiedenen* Anwendungsbereichen, Hinweisen und Hilfestellungen sowie Dokumentation von häufig gemachten Fehlern bei der Anwendung der Standards, Problemdarstellung und Aufnahme von Lösungsansätzen bei Problemen bezgl. der Anwendung der Standards. Je ca. 15-mal wird eine Anpassung auf verschiedene Kontexte oder Evaluationsarten (z. B. Selbstevaluation) gewünscht. Konkrete Anregungen zu bestimmten Standards sind selten. Auch die anderen Antworten streuen stark auf viele Unterkategorien mit jeweils wenigen Nennungen.

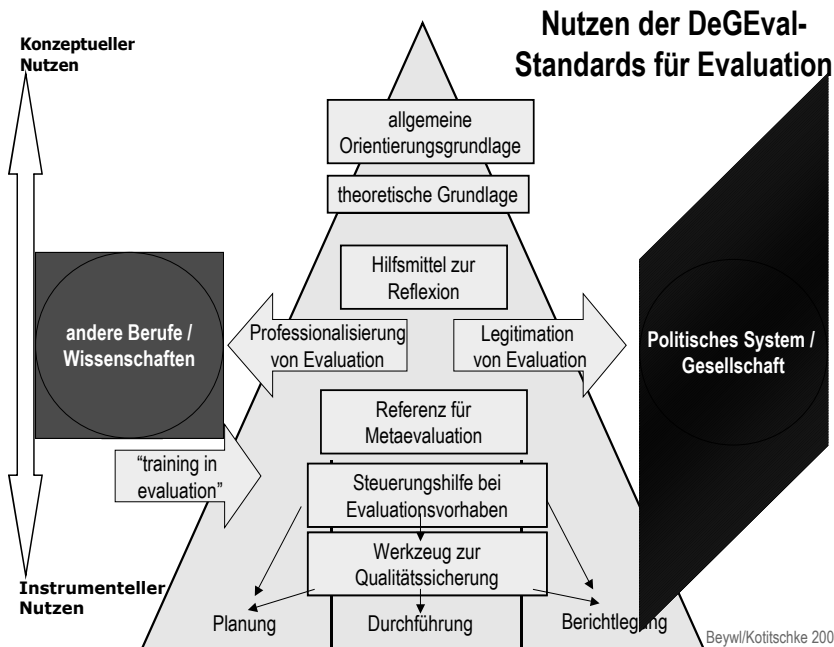


Abbildung 2: Nutzung der DeGEval-Standards für Evaluation (Quelle: Beywl & Kotitschke, 2004b, S. 8)

Schließlich konnten in der Auswertung einige Spannungsthemen herausgearbeitet werden, namentlich:

1. Maximalstandards versus Minimalstandards:
Also eher Orientierungsfunktion oder eher Setzen einer unteren, einzuhaltenden Qualitätsschwelle
2. Knappe Standards versus detaillierte Standards:
Also möglichst kurze Standardtexte oder ausführlich beschriebene und erläuterte Standards
3. Feldspezifische Standards versus feldunspezifische Standards:
Also ob die Standards auf unterschiedliche Anwendungsbereiche anzupassen bzw. zuzuschneiden sind oder wie bisher feldübergreifend sein sollen.

Zusammenfassend ist festzuhalten: Die Antworten in der Onlinebefragung zur Steuerung des Revisionsprozesses zeugen von einem starken Interesse an den Standards und ihrer weiteren Entwicklung. Dies kommt auch in zahlreichen Verbesserungsvorschlägen und Kommentaren zum Ausdruck. Auch die Mitarbeitsbereitschaft an einem Revisionsprozess ist ausgeprägt.

Der DeGEval-Vorstand verfolgte im Anschluss an diese Befragung weiterhin das Ziel, eine umfassende Revision mit dem Ziel einer deutlichen Erweiterung in Richtung eines Handbuches der DeGEval-Standards. Es gelang jedoch trotz mehrerer Anläufe lange nicht, die hierfür erforderlichen personellen und finanziellen Ressourcen zu erschließen.

5.2 Die erste Revision 2012–2016

Mit der einstimmig erfolgten Zustimmung der Mitgliederversammlung der DeGEval am 4. Oktober 2001 war auch der Auftrag an den Vorstand verbunden, die Standards für Evaluation zusammen mit den von der Kommission erarbeiteten Erläuterungen zu verbreiten und in die fachliche und öffentliche Diskussion einzubringen. Die Standards waren seitdem wichtiger Bezugspunkt für mit Evaluation befasste Personen und Organisationen. Die „Standards der Evaluation“ sind das bedeutendste inhaltliche Produkt der DeGEval. Und dies gilt nicht nur im Hinblick auf die verbreitete Verständigung über die Qualität von Evaluation. Auch für die Etablierung der „Gesellschaft für Evaluation“ waren die Standards bedeutend.

Dennoch: Mit ihrem Erfolg wurde aus der Mitgliedschaft der DeGEval auch zunehmend die Frage angesprochen, ob und ggf. wie die Standards der Evaluation revidiert werden sollten. Diese Revision war, wie gesagt, bereits in ihrer ersten Fassung vorgesehen. Im Jahr 2012 wurden Personen, die am ersten Prozess der Standard-Entwicklung beteiligt waren und diejenigen eingeladen, die bei der letzten Jahrestagung Interesse am Thema bekundet hatten, die Frage einer möglichen Revision zu diskutieren. Die Sichtung der Standards erbrachte einen moderaten Änderungsbedarf, der sich u. a. auf folgende Punkte bezog:

- Ergänzung durch Operationalisierung,
- Verbesserung der Verständlichkeit,
- Verringerung von quantitativen Disproportionalitäten zwischen den Standardgruppen,
- stärkere Berücksichtigung der Multiperspektivität einer Evaluation,
- Beseitigung bzw. Moderation von Widersprüchen,
- Abgrenzung zu verwandten Begriffen und Konzepten, z. B. Monitoring, Audit etc.

Mit der Revision, so eine grundsätzliche Überlegung, könnte eine Brücke zur Anwendung geschlagen werden: „Von der Orientierung zur Handlungsanleitung“. Der Mitgliederversammlung sollte ein Vorschlag für die Durchführung eines Revisionsprozesses und die Bildung einer Ad-hoc-Gruppe vorgelegt werden. Auf der Mitgliederversammlung 2012 sprachen sich die anwesenden Mitglieder einstimmig hierfür aus: „Die Mitgliederversammlung beauftragt den Vorsitzenden der DeGEval mit der Gründung einer Ad-hoc-Gruppe ‚Revision der Standards der Evaluation‘. Diese Gruppe hat den Auftrag, die bisherigen Standards zu reflektieren und bis spätestens zur Jahrestagung 2014 der Mitgliederversammlung in Zürich einen Vorschlag für (ggf. revidierte) Standards zur Diskussion vorzulegen. Die Ad-hoc-Gruppe ‚Standards‘ öffnet ihre Arbeit durch Einbeziehung eines Kreises von interessierten Mitgliedern, die sich zur Kommentierung von Zwischenergebnissen verpflichten. Zusätzlich werden Zwischenergebnisse den Mitgliedern auf der Homepage zur Kenntnis gegeben und Einwirkungsmöglichkeiten geschaffen. Die Ad-hoc-Gruppe prüft insbesondere auch, inwieweit die Verzahnung von Standards und ethischen Vorgaben (Code of Conduct) sinnvoll ist oder ob zwei getrennte Dokumente erstellt und abgestimmt werden müssen.“

Hiermit begann ein intensiver Diskussionsprozess innerhalb der Arbeitsgruppe. Die Mitgliedschaft wurde anlässlich der Mitgliederversammlungen über den Verlauf informiert. Im Kontext der Jahrestagung 2013 fand eine Session statt, bei der die notwendigen weiteren Prozesse und Inhalte diskutiert wurden. Im Winter 2013 fand in Bern (CH) ein von EvalPartners geförderter Workshop mit der Schweizerischen Evaluationsgesellschaft SEVAL statt, die parallel an der Revision ihrer Standards arbeitete. Hier hat sich im gemeinsamen Austausch und nach Beiträgen eingeladenen Expertinnen und Experten bei der DeGEval der Ansatz gefestigt, die Standards lediglich in Details, also „konservativ“ zu revidieren.

Vor der Jahrestagung 2015 legt die AG Standards den Mitgliedern der DeGEval einen Entwurf der Revision vor. Die Sprecherinnen und Sprecher der Arbeitskreise wurden gebeten, eingehende Kommentierungen und Vorschläge in eine endgültige Konzeption einzuspeisen und mit der AG Standards und dem Vorstand der DeGEval abzustimmen. Auf der Jahrestagung fand zudem eine Session statt, in der es um die Sammlung von Ideen zu der Frage ging, wie die Praxisrelevanz der Standards gesteigert werden kann. Auch wurde diskutiert, wie sie intern und – vor allem – extern kommuniziert werden sollen, damit sie die intendierten positiven Wirkungen für die Qualität von Evaluation entfalten können. Im Prozess der Revision war es wichtig, die Mitglieder der DeGEval zu beteiligen.

Zusammenfassend kann festgestellt werden, dass es sich als sinnvoll erwiesen hat, die erste Version der Standards der Evaluation zu begutachten und zu revidieren. Die Qualität der ursprünglichen Arbeit wird dadurch deutlich, dass die Revision inkrementell erfolgen konnte. Bei der Überarbeitung der Standards ergaben sich keine Gründe für eine grundsätzliche Neufassung.

Die Revisionsgruppe nahm nur moderate Änderungen an den Standards vor (vgl. zum Folgenden Hense, Böttcher, Caspari, Kalman & Meyer, 2018). Vor allem die elegante Grundstruktur mit den vier griffigen Qualitätsmerkmalen „nützlich, durchführbar, fair und genau“ blieb erhalten. Die Ergänzung eines fünften Bereichs „Evaluationsverantwortlichkeit“, wie sie das Joint Committee vorgenommen hatte, wurde als inhaltlich nicht begründet verworfen. Die Option einer Umstellung auf eine prozedurale Gliederung, in der Standards je Phase eines Evaluationsprojekts definiert werden, wurde v. a. aus drei Gründen verworfen. Erstens bedingt eine solche Gliederung immer Vorannahmen über ein Evaluationsprojekt, die nicht in allen Fällen zutreffen. Dies widerspräche dem generischen Anspruch der Standards, für alle Varianten, Konstellationen und Einsatzfelder der Evaluation zu gelten. Zweitens sollte die etablierte ‚Marke‘ der vier griffigen Standardbereiche nicht leichtfertig aufgegeben werden. Drittens besteht eine wichtige Funktion der Standards nicht nur darin, Orientierung für die Praxis zu geben, was vielleicht mit einer prozeduralen Gliederung leichter sein könnte, sondern auch Kriterien für die Meta-Evaluation bereitzustellen, also für die Beurteilung von Güte und Nutzen von Evaluation selbst. Die Revisionsgruppe hielt das nur in der gegebenen Struktur für möglich. Als Kompromiss enthält auch die revidierte Fassung ein funktionales Inhaltsverzeichnis, das für die Anwendung der Standards in verschiedenen Evaluationsphasen Orientierung geben soll.

Inhaltliche Änderungen wurden somit eher auf Detailebene vorgenommen. Dabei reagieren viele Überarbeitungen und Ergänzungen auf fachliche Entwicklungen

der jüngeren Zeit. Dazu gehört etwa die gewachsene Bedeutung von Wirkmodellen und Programmtheorien in der Evaluation, auf deren Rolle jetzt in verschiedenen Einzelstandards klarer verwiesen wird. Ähnliches gilt für den Datenschutz, der an relevanten Stellen stärker betont wurde. Ein drittes Beispiel ist das Thema Implementationsstreue, das aus guten Gründen in jüngerer Zeit vermehrt Aufmerksamkeit in der Evaluationsliteratur erfahren hat.

Neben diesen inhaltlichen Aktualisierungen und Verbesserungen ‚unter der Motorhaube‘ wurde viel Aufwand in eine leichtere Verständlichkeit und Zugänglichkeit investiert. Sie sollen vor allem durch eine klarere und konsistenter umgesetzte Strukturierung der Erläuterungen zu den Einzelstandards erreicht werden. Zusätzlich wurden die konzeptionellen und begrifflichen Grundlagen in einem eigenen Abschnitt klarer herausgestellt. Eine weitere wichtige Weiterentwicklung war die erstmalige Ergänzung eines Glossars der Standards für Evaluation, das für ein konsistentes Begriffsverständnis, vielleicht auch über die Standards hinaus, sorgen sollte. Die revidierten Standards befinden sich zum Download auf der Homepage der DeGEval, ebenso wie Angaben zum Bezug der Druckausgabe.

6. Potenzial der Standards

Die DeGEval-Standards sind Grundlage für professionelle Evaluation. Sie beschreiben die zentralen Prinzipien, die aus Sicht der DeGEval zwingend zu beachten sind. Der Gesamtheit der Standards einer professionellen Evaluation ist hingegen in der Praxis womöglich nicht immer adäquat und umfassend zu entsprechen. Die jeweiligen Standards haben auch unterschiedliche und je nach konkreter Evaluation unterschiedlich gewichtete Relevanz. Aber für eine gute Evaluation gilt, dass sich Evaluatorinnen und Evaluatoren in ihrer Arbeit auf die Standards beziehen und erläutern, ob und inwieweit die darin formulierten Anforderungen Eingang in ihre Arbeit gefunden haben.

Von Auftraggebenden einer Evaluation wird erwartet, dass auch sie sich bei Auftragsbeschreibung und -vergabe an den Standards orientieren. Die Standards als gemeinsame Orientierung befördern die Professionalität, mithin die Güte und den Nutzen der Evaluation. Auch für den Fall, dass Personen oder Organisationen ihre eigene Arbeit eigenverantwortlich evaluieren wollen, liefern die Standards eine Orientierung. Die DeGEval hat zwar auch weitere und weiterführende Publikationen für Auftraggebende und für die Selbstevaluation erarbeitet, aber auch in deren Zentrum stehen die Standards. Insofern sind Standards für alle Akteure leitend, die mit Evaluation zu tun haben.

Mit den „Standards für Evaluation“ wollen die in der *DeGEval – Gesellschaft für Evaluation* organisierten Personen und Organisationen einen Orientierungsrahmen zur Bestimmung einer „guten“ Evaluation geben. Mit den Standards wurde versucht, die Idee einer professionellen Evaluation zu konkretisieren. Mit den „Standards für Evaluation“ wird das Konzept Evaluation gewissermaßen induktiv definiert: Aus den Standards lässt sich ablesen, was Evaluation ausmacht. Dieses Verständnis erlaubt auch, die Differenzen zwischen anderen bewertenden Verfahren wie z. B. Feedbacks oder Audits zu verdeutlichen wie auch die Differenzen zu einer Wissenschaft zu markieren, die sich letztlich lediglich in der eigenen Community rechtfertigen muss.

Mit der Revision wurde der ausdrückliche Wunsch verbunden, diese Neuerungen einer Prüfung auf Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit zu unterziehen.

Literatur

- Abbott, A. (1991). The Order of Professionalization. *Work and Occupations*, 18(4), 355–384. doi: <https://doi.org/10.1177/0730888491018004001>
- Altschuld, J. W. (1999). The case for a voluntary system for credentialing evaluators. *American Journal of Evaluation*, 20(3), 507–517.
- American Psychological Association (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington D.C.: APA.
- American Psychological Association (1974). *Standards for Educational and Psychological Tests and Manuals*. Rev. ed. Washington D.C.: APA.
- Balzer, L. (2013). Bekanntheitsgrad und Nutzung der Evaluations-Standards der Schweizerischen Evaluationsgesellschaft (SEVAL) sowie erste Modifikationshinweise: Ergebnisse der Mitgliederbefragung. *LeGes – Gesetzgebung & Evaluation*, 24(2), 439–458.
- Balzer, L. & Beywl, W. (2018). *evaluiert. – erweitertes Planungsbuch für Evaluationen im Bildungsbereich* (2. überarb. Aufl.). Bern: h.e.p.
- Beywl, W. (1988). *Zur Weiterentwicklung der Evaluationsmethodologie. Grundlegung, Konzeption und Anwendung eines Modells der responsiven Evaluation*. Frankfurt: Peter Lang.
- Beywl, W. (1996). Anerkannte „Standards“ und „Leitprinzipien“ der amerikanischen Evaluation. In M. Heiner (Hrsg.), *Qualitätsentwicklung durch Evaluation* (S. 85–107). Freiburg i. Br.: Lambertus.
- Beywl, W. (1999). Standards für Evaluationen. *DeGEval... News*, 2, 1–4.
- Beywl, W., Bestvater, H. & Friedrich, V. (2011). *Selbstevaluation in der Lehre. Ein Wegweiser für sichtbares Lernen und besseres Lehren*. Münster: Waxmann.
- Beywl, W. & Kotitschke, H. (2004a). *Befragung zur Steuerung des Revisionsprozesses der DeGEval-Standards für Evaluation (REVAL) – Bericht zur quantitativen Datenauswertung*. Verfügbar unter http://www.degeval.de/fileadmin/DeGEval-Standards/REVAL_Bericht_quantitativ_VER_Feb_05.pdf
- Beywl, W. & Kotitschke, H. (2004b). *Befragung zur Steuerung des Revisionsprozesses der DeGEval-Standards für Evaluation (REVAL) – Bericht zur qualitativen Datenauswertung*. Verfügbar unter http://www.degeval.de/fileadmin/DeGEval-Standards/REVAL_qualBericht_Ver_Feb_05.pdf
- Beywl, W., Künzli, C., Messmer, R. & Streit, C. (2015). Forschungsverständnis Pädagogischer Hochschulen – ein Diskussionsbeitrag. *Beiträge zur Lehrerinnen- und Lehrerbildung*, 33(1), 134–151.
- Beywl, W. & Widmer, T. (2006). Die ‚Standards‘ im Vergleich mit weiteren Regelwerken zur Qualität fachlicher Leistungserstellung. In Joint Committee on Standards for Educational Evaluation & J. R. Sanders (Hrsg.), *Handbuch der Evaluationsstandards: Die Standards des „Joint Committee on Standards for Educational Evaluation“* (Übers. und für die dt. Ausg. erw. von Wolfgang Beywl und Thomas Widmer) (3. Aufl., S. 263–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Beywl, W. & Widmer, T. (2012). Die Standards für Programmevaluation des Joint Committee on Standards for Educational Evaluation. „Übersetzung der Kurzfassung aus: Yarbrough, D. B., Shulha, L. M., Hopson, R. K. & Caruthers, F. A. (2011). *The program evaluation standards. A guide for evaluators and evaluation users* (3rd. ed.). Thousand

- Oaks: Sage.“ Aarau und Zürich: Pädagogische Hochschule FHNW/Universität Zürich. doi: <https://doi.org/10.5167/uzh-63498>
- Böttcher, W. & Hense, J. (2015). Professionelle Evaluation oder Evaluation als Profession? In V. Hennefeld, W. Meyer & S. Silvestrini (Hrsg.), *Nachhaltige Evaluation?: Auftragsforschung zwischen Praxis und Wissenschaft* (S. 101–120). Münster: Waxmann.
- Curnow, C. K. & McGonigle, T. P. (2006). The effects of government initiatives on the professionalization of occupations. *Human Resource Management Review*, 16(3), 284–293. doi: <https://doi.org/10.1016/j.hrmmr.2006.06.001>
- DeGEval – Gesellschaft für Evaluation (2002). *Standards für Evaluation*. Köln: DeGEval.
- Evaluation Research Society Standards Committee (1982). Evaluation Research Society standards for program evaluation. In P. H. Rossi (Hrsg.), *Standards of evaluation practice. New Directions for Program Evaluation* 15 (S. 7–19). San Francisco: Jossey-Bass.
- Evaluation Research Society Standards Committee (1990). Standards der Evaluation der Evaluation Research Society. In U. Koch & W. W. Wittmann (Hrsg.), *Evaluationsforschung* (S. 311–315). Berlin: Springer.
- Fox, C. J. (1992). What do we Mean When we Say “Professionalism?”: A Language Usage Analysis for Public Administration. *The American Review of Public Administration*, 22(1), 1–17. doi: <https://doi.org/10.1177/027507409202200101>
- Frankel, M. S. (1989). Professional Codes Why, How, and with What Impact? *Journal of Business Ethics*, 8(2/3), 109–115. doi: <https://doi.org/10.1007/BF00382575>
- Guba, E. G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park: Sage.
- Hellstern, G. M. & Wollmann, H. (Hrsg.) (1984). *Handbuch zur Evaluierungsforschung*. Opladen: Westdeutscher Verlag.
- Hense, J., Böttcher, W., Caspari, A., Kalman, M. & Meyer, W. (2018). Revision der Standards für Evaluation. *Zeitschrift für Evaluation*, 17(1), 174–176.
- Hilton, S. & Southgate, L. (2007). Professionalism in medical education. *Teaching and Teacher Education*, 23(3), 265–279. doi: <https://doi.org/10.1016/j.tate.2006.12.024>
- Joint Committee on Standards for Educational Evaluation (1981). *Standards for evaluations of educational programs, projects, and materials*. New York: McGraw-Hill.
- Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2. Aufl.). Thousand Oaks: Sage.
- Joint Committee on Standards for Educational Evaluation (2003). *The student evaluation standards: How to improve evaluations of students*. Thousand Oaks: Corwin Press.
- Joint Committee on Standards for Educational Evaluation (2009). *The personnel evaluation standards: How to assess systems for evaluating educators*. Thousand Oaks: Corwin Press.
- Joint Committee on Standards for Educational Evaluation (2011). *The program evaluation standards: A guide for evaluators and evaluation users*. Los Angeles: Sage.
- Kuji-Shikatani, K. (2015). Credentialed Evaluator Designation Program, the Canadian Experience. *New Directions for Evaluation*, 2015(145), 71–85. doi: <https://doi.org/10.1002/ev.20112>
- Lincoln, Y. S. (1985). The ERS standards for program evaluation: Guidance for a fledging profession. *Evaluation and Program Planning*, 8(3), 251–253.
- Love, A. & Russon, C. (2005). Evaluation standards in an international context. *New Directions for Evaluation*, 2004(104), 5–14.
- Madaus, G. F. & Stufflebeam, D. L. (2000). Program evaluation: A historical overview. In D. L. Stufflebeam, G. F. Madaus & T. Kellaghan (Hrsg.), *Evaluation models. Viewpoints on educational and human services evaluation* (S. 3–18). Boston: Kluwer.

- Mieg, H. A. (2010). Professionalisierung: eine dreifache Autonomieregulation. In K. Gräfin von Schlieffen, R. Breinlinger, F. Dauner & R. Greger (Hrsg.), *Professionalisierung und Meditation* (S. 15–27). München: C. H. Beck.
- Molander, E. A. (1987). A Paradigm for Design, Promulgation and Enforcement of Ethical Codes. *Journal of Business Ethics*, 6(8), 619–631. doi: <https://doi.org/10.1007/BF00705778>
- Newman, D. L. & Brown, R. D. (1996). *Applied ethics for program evaluation*. Thousand Oaks: Sage. Verfügbar unter: <http://www.loc.gov/catdir/enhancements/fy0655/95036416-d.html>
- Newman, D. L., Scheirer, M. A., Shadish, W. R. & Wye, C. (1995). Guiding Principles for Evaluators. *New Directions for Program Evaluation*, 1995(66), 19–26.
- Rouge, J. C. (2005). The origin and development of the African evaluation guidelines. *New Directions for Evaluation*, 2004(104), 55–66.
- Russon, C. (2005). Cross-cutting issues in international standards development. *New Directions for Evaluation*, 104, 89–93.
- Sanders, J. R. (2006). Aspekte der Entwicklung und Verbreitung der Evaluationsstandards. In Joint Committee on Standards for Educational Evaluation & J. R. Sanders (Hrsg.), *Handbuch der Evaluationsstandards: Die Standards des „Joint Committee on Standards for Educational Evaluation“* (S. 241–245). Wiesbaden: VS-Verlag.
- Scriven, M. (1969). An Introduction to Meta-Evaluation. *Educ Prod Rep.*, 2, 36–38.
- SEVAL – Schweizerische Evaluationsgesellschaft (2016). *Evaluationsstandards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards)*. Bern: SEVAL.
- Widmer, T. (1996). *Meta-Evaluation: Kriterien zur Bewertung von Evaluationen*. Zugl.: Zürich, Univ., Diss., 1996. Bern: Haupt.
- Widmer, T. (2005). The development and status of evaluation standards in Western Europe. *New Directions for Evaluation*, 104, 31–42.
- Widmer, T. (2006). Evaluationsstandards in Europa. In Joint Committee on Standards for Educational Evaluation (Hrsg.), *Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation“* (S. 301–312). Wiesbaden: VS Verlag.
- Widmer, T. (2011). Zehn Jahre Evaluationsstandards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards). *Schweizerische Zeitschrift für Kriminologie*, 10(2), 23–30.
- Widmer, T. (2015). Développements et perspectives entre profession et discipline. Evaluation quo vadis? In K. Horber-Papazian (Hrsg.), *Regards croisés sur l'évaluation en Suisse*. Lausanne: Presses polytechniques et universitaires romandes.
- Widmer, T. (2017). Geschichte der Evaluation im schweizerischen politischen System. In F. Sager, T. Widmer & A. Balthasar (Hrsg.), *Evaluation im politischen System der Schweiz* (S. 51–66). Zürich: NZZ Libro.
- Widmer, T. & Beywl, W. (2006). Die Übertragbarkeit der Evaluationsstandards auf unterschiedliche Anwendungsfelder. In Joint Committee on Standards for Educational Evaluation (Hrsg.), *Handbuch der Evaluationsstandards* (S. 247–261). Wiesbaden: VS Verlag.
- Widmer, T., Landert, C. & Bachmann, N. (2000). *Evaluations-Standards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards)*. Verfügbar unter: <https://www.zora.uzh.ch/id/eprint/164982/>
- Widmer, T., Rothmayr, C. & Serdült, U. (1996). *Kurz und gut? Qualität und Effizienz von Kurzevaluationen*. Zürich: Rüegger.
- Wilensky, H. L. (1964). The Professionalization of Everyone? *The American Journal of Sociology*, 70(2), 137–158.
- Yarbrough, D. B., Shulha, L. M. & Caruthers, F. (2005). Background and history of the Joint Committee's program evaluation standards. *New Directions for Evaluation*, 104, 15–30.