Standards in der Evaluation beruflicher Bildung: Anspruch und Wirklichkeit

Frühjahrstagung des Arbeitskreises Berufliche Bildung im Mai 2014

Michael Kalman[1]

Am 16. Mai 2014 fand die Frühjahrstagung des Arbeitskreises Berufliche Bildung an der Humboldt Universität (HU) zu Berlin statt. Knapp 20 Expertinnen und Experten tauschten sich zum Thema „Standards in der Evaluation beruflicher Bildung: Anspruch und Wirklichkeit" aus.

Das Thema ist aktuell, befindet sich die Gesellschaft für Evaluation e.V. (DeGEval) doch in einem Revisionsprozess ihrer Standards, die inzwischen ein wenig ‚in die Jahre gekommen' sind. Im Oktober 2001 wurden die „Standards für Evaluation" von der Mitgliederversammlung einstimmig angenommen, 2002 erfolgte dann die Veröffentlichung. Obwohl die Revision seinerzeit schon mitgedacht worden war, dauerte es doch 11 Jahre bis zum Mitgliederbeschluss von 2013, bis der Revisionsprozess tatsächlich angestoßen wurde. Die Tagung verstand sich so auch als Inputgeber für die Revision.

Diese Inputs sollten aus der Praxis kommen. Daher wurden die Referent(inn)en Rüdiger Preißer und Tülin Engin gebeten, bereits durchgeführte und abgeschlossene Evaluationsprojekte vorzustellen und retrospektiv hinsichtlich der Relevanz und des Erfüllungsgrades der Standards für Evaluation zu bewerten.

Michael Kalman gab zu Beginn eine kurze Einstimmung in den „generischen Anforderungskatalog" der 25 Standards, die in die vier Gruppen Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit aufgeteilt sind, und schlussfolgerte, dass „der generische Charakter der Standards […] mitunter hohe Herausforderungen an die Evaluierenden" stellt, „um die Anforderungen auf ihr konkretes Evaluationsprojekt zu beziehen." Schließlich beanspruchen die Standards Geltung für viele, wenn nicht für alle Felder, Akteure, Zwecke, Evaluationsgegenstände und Evaluationsarten. Und: Sie schreiben keine bestimmte Methode bzw. einen spezifischen Methodenmix vor.

1 Kalman Consult, Berlin

Zur Adressierung dieser Herausforderungen stellt die Standards-Broschüre Interpretations- und Anwendungshilfen bereit, nämlich Einleitung/Vorwort, Standards-Erläuterungstexte, ein funktionales Inhaltsverzeichnis und eine Transformationstabelle zu den US-amerikanischen Joint Committee on Standards und den Standards der Schweizerischen Evaluationsgesellschaft (SEVAL). Die Brückenfunktion zur Evaluationspraxis dieser Interpretations- und Anwendungshilfen – so verdienstvoll sie sind – wurde kritisch hinterfragt. Die zuweilen schwierige Aufgabe der Abwägung bei der Auslegung und Verwendung der Standards kann dem Evaluationsteam freilich nicht abgenommen werden. Dies umso mehr, als sich einige Standards durchaus widersprechen können, z.B. aus den Gruppen Durchführbarkeit und Genauigkeit. Es wird stets angesichts begrenzter Ressourcen und bezogen auf Evaluationsgegenstand und Fragestellung abzuwägen sein, wie genau das Evaluationsteam sein muss, um auf der einen Seite die Gütekriterien der empirischen Sozialforschung zu erfüllen und auf der anderen Seite die Durchführbarkeit des Evaluationsprojekts angesichts häufig knapper Ressourcen nicht zu gefährden. Auch ist zu fragen, ob in jedem Fall eine Evidenzbasierung erforderlich ist, um die Nützlichkeit der Evaluationsergebnisse für den Auftraggeber sicherzustellen.

Die Entlastung in diesem schwierigen Geschäft der Abwägung und Auslegung ist gleichwohl in Sicht, weil die Nichtanwendung bestimmter Standards ausdrücklich möglich ist; auch darf und muss die Art und Weise der Erfüllung von Standards variieren und schließlich: Die Standards sind aktuell keine Grundlage für Dienstleistungszertifizierungen, was jedoch auch den Vorteil hat, dass kein verengender Formalismus vorliegt.

Insgesamt sah Kalman in der Ableitung von handlungsleitenden Checklisten aus dem funktionalen Inhaltsverzeichnis eine Möglichkeit, die „Brücke zur Praxis“ flankierend zu den Standards zu stärken.

Für die Tagung wurden zwei Perspektiven aufgezeigt, nämlich die Überprüfung des praktischen Umgangs mit den Standards und die Überprüfung der Standards in ihrer Relevanz für die Praxis (Revision). Den beiden Referent(inn)en wurden die folgenden Fragen an die Hand gegeben:

- Hat der Auftraggeber die Berücksichtigung der Standards vorgegeben?
- Wenn ja, wie hat der Auftraggeber die Einhaltung der Standards kommuniziert/kontrolliert?
- Welche Relevanz/welchen Stellenwert hatten die Standards für die verschiedenen Projektphasen (Konzept, Einstiegsphase, Durchführungsphase, Berichterstattung, Nachprojektphase)?
- Welche expliziten und impliziten Rekurse auf die Standards wurden im Evaluationsprojekt vorgenommen?
- In der Retrospektive: Welche Standards wurden erfüllt, welche nicht und warum?

Der Erziehungs- und Sozialwissenschaftler *Rüdiger Preißer* präsentierte „Schlaglichter auf die Standards für Evaluation aus der Perspektive ausgewählter Evaluationsprojekte“. Preißer stellte folgende Projekte vor:

- Qualität der Lehre und Studienbedingungen an der HU Berlin (1990-93);
- ProfilPass (DIE, DIPF, 2004/05);

– Berufsvorbereitende Bildungsmaßnahmen (BvB, 2009/10).

Bereits zu Beginn stellte er klar, dass er Standards für Evaluation nie zur Grundlage genommen habe. Stattdessen waren für ihn die Regeln der empirischen Sozialforschung handlungsleitend.

Im Rahmen des Projekts Qualität der Lehre und Studienbedingungen an der HU Berlin (1990-93) wurden Studierende aus 14 Studiengängen mit insgesamt ca. 8000 Bewertungsbögen zu ihrer Lehrveranstaltung befragt. Auf diese Weise wurden ca. 300 Lehrveranstaltungen (LV) bewertet. Zu den erhobenen Merkmalen gehörte der Input (Infrastruktur, Rahmenbedingungen der LV), der Prozess (Durchführung) und der Outcome (Lernfortschritt, Studienverhalten). Aus den Befragungsergebnissen wurden verschiedene Auswertungen vorgenommen, z.B. zu Strukturvergleichen zwischen LV-Typen, zu Studiengängen, Studienabschnitten und Studierenden-Typen im Sinne soziodemografischer Merkmale. Ferner wurden die Bewertungsdimensionen Input und Durchführung sowie der Outcome im Sinne des subjektiven Lernerfolgs ermittelt.

Preißer setzte sich kritisch mit dem Untersuchungsdesign auseinander – zunächst immanent. So kritisierte er die zu starke Konzentration auf die Outcome-Dimension bei Vernachlässigung des Inputs. So hätte weder ein Curriculum noch ein didaktisches Konzept existiert; die Qualifikation der Lehrenden sei ebenfalls nicht vorhanden gewesen. Zudem sei die Outcome-Dimension nur aus der Zielgruppenperspektive im Sinne des subjektiv bewerteten Lernerfolgs erhoben worden; dasselbe gelte für die Prozessqualität, wo „veraltete Indikatoren“ verwendet worden seien.

Auf einer grundsätzlicheren Ebene lautete seine Kritik: „Wieso konnte das Ergebnis nicht lauten, dass Lehre prinzipiell falsch angelegt ist: weitgehend dem traditionellen Vermittlungsmodell von Wissen folgt, anstatt auf Grundlage konstruktivistischer Lerntheorie („Shift from Teaching to learning“) zu operieren?“ Sein skeptischer Dreiklang bezogen auf die Standards für Evaluation – die es freilich im deutschsprachigen Raum in den 1990er Jahren so noch nicht gegeben hatte:

– Sind die Standards verwendet worden?
– Hätten die Standards etwas geändert (am Resultat, an der Durchführung)?
– Hätten andere Standards etwas geändert?

Der Referent sprach damit die Steuerungsmöglichkeiten von Standards in Bezug auf die Qualität konkreter Evaluationsprojekte an. Insgesamt – so kann implizit aus dem Referat von Preißer geschlossen werden – würde die Anwendung der Standards nicht bewirkt haben, dass im Evaluationsprojekt die Input-Dimension stärker beachtet oder die Grundanlage des Evaluationsgegenstandes prinzipiell hinterfragt worden wäre.

In einem weiteren Projekt, die Evaluation des ProfilPasses im Jahr 2004/05, sollten Entscheidungsgrundlagen über die bundesweite Einführung dieses Instrumentes erbracht werden. Zudem sollten Anhaltspunkte für Modifizierungen und Verbesserungen generiert werden. Der ProfilPass ist ein Portfolio-Instrument zur Exploration und Dokumentation der individuellen Kompetenzen. Es wurde im Rahmen der Evaluation eine standardisierte schriftliche Befragung der Nutzer(innen), teilstandardisierte schriftliche Befragung der Teilnehmenden an der Qualifizierung zur ProfilPass-Beratung und leitfadengestützte Interviews mit Berater(inne)n durchge-

führt. Die Evaluation erbrachte verschiedene Ergebnisse zur Outcome-Dimension insgesamt und für Personengruppen, ermöglichte z.B. Strukturvergleiche zwischen Outcome-Dimensionen und Input-/Durchführungs-Dimensionen und gewährleistete Hypothesengenerierung aufgrund der Bewertungen der Berater(innen). Für Preißer war der Evaluationsgegenstand, ein „übliches" Verfahren, „richtig" entwickelt und „richtig" evaluiert; aber wenn das Verfahren losgelöst von Zielen eingesetzt werde, drohe eine Verselbstständigung. Der Referent bezweifelte zudem, ob Schlussfolgerungen aus der Evaluation für die weitere Entwicklung/die endgültige Einführung des ProfilPasses gezogen wurden. Letzteres wird ja durch den Standard N8 „Nutzung und Nutzen der Evaluation" angesprochen.

Im Jahr 2009/10 evaluierte Preißer Berufsvorbereitende Bildungsmaßnahmen (BvB) durch die Erfassung (Rating) der „erworbenen Kompetenzen" bei rund 300 Jugendlichen (‚Wirkungskontrolle'). Dabei erfolgten kein Vergleich mit den Anfangskompetenzen und auch keine direkte Befragung der Zielgruppe. Zusätzlich wurden Fallstudien mit Experteninterviews und Dokumentenanalysen bei 10 Bildungsträgern zu Maßnahmenkonzeption, -durchführung und -resultaten durchgeführt. Offiziell sollte im Rahmen einer ‚Wirkungskontrolle' evaluiert werden, ob es bei den Teilnehmer(inne)n einen Kompetenzzuwachs gegeben hat und ob ein Ausgleich der Kompetenzdefizite erfolgte. Inoffiziell ging es auch um die Input- und Durchführungsqualität. Preißer warf in diesem Zusammenhang das Verhältnis von Wirkungsevaluation zur Input-Evaluation auf und konstatierte, dass eine Wirkungskontrolle, die den Namen verdient, in diesem Projekt nicht möglich war. Er bemängelte, dass es kaum Evaluationen des Übergangssystems gebe und zeigte sich skeptisch, ob bisherige Evaluationen irgendetwas bewirkt haben.

Insgesamt beschlich den Referenten bei allen drei Projekten ein Unwohlsein, obwohl diese Vorhaben methodisch korrekt durchgeführt wurden. Und er fragte sich, inwieweit die Anwendung von Standards diese Projekte auf ein ihm angemessen erscheinendes Qualitätslevel hätten heben können.

Anhand seiner Projekte legte Preißer dar, dass der Zuschnitt des Evaluationsgegenstandes unzureichend oder gar falsch sein kann – etwa, wenn die Input-Dimension ausgenommen wird. Der dafür vorgesehene Standard G 1 „Beschreibung des Evaluationsgegenstandes" – so möchte der Verfasser hinzufügen – entlässt das Evaluationsteam ja nicht aus der Verantwortung, den Evaluationsgegenstand entsprechend der selbst aufgestellten oder vorgegebenen Kriterien zu bestimmen. Dieser Standard geht jedoch nicht so weit, dass er Detailvorgaben macht – etwa im Sinne: Die Input-Dimension müsse bei jedem Evaluationsgegenstand mit bedacht werden.

Preißer bemängelte, dass das Schema Input – Aktivitäten – Output – Outcome, wie es etwa der Programmbaum der Firma Univation impliziere, so von den Standards für Evaluation nicht nahegelegt werde. Aus seiner Sicht seien die Standards auch zu vage formuliert. Er empfahl, die Standards besser zu operationalisieren. Ferner sollte man Mindeststandards/Gütekriterien formulieren. Dabei sollte man sich vom Prinzip der Gleichgewichtigkeit der Standards verabschieden. Insgesamt unterstütze er aber eher eine Initiative für die Erarbeitung von Grundlagen für bessere Evaluationskonzeptionen anstatt die Revision der Standards.

In der Summe diskutierte Preißer hochrelevante Aspekte, welche die Qualität von konkreten Evaluationsprojekten betreffen, die aber durch das gröbere Raster der Standards teilweise durchzufallen scheinen.

Die Psychologin *Tülin Engin* vom uzbonn – Gesellschaft für empirische Sozialforschung und Evaluation referierte zu „Nutzung und Nutzen der DeGEval-Standards zur Meta-Evaluation – Ein Erfahrungsbericht". Frau Engin stellte dabei eine vom Bundesinstitut für Berufsbildung (BIBB) in Bonn beauftragte Meta-Evaluation vor. Dabei sollte eine abgeschlossene Evaluation noch einmal bewertet werden. Letztendlich handelte es sich um eine kriteriengeleitete Dokumentenanalyse, denn die Informationsbasis für die Meta-Evaluation bestand aus dem Abschlussbericht der Evaluation und dem entsprechenden Kurzbericht sowie der Leistungsbeschreibung der damaligen Ausschreibung des BIBB. Diese Informationsbasis wurde ergänzt durch Abstimmungsgespräche mit dem BIBB als Auftraggeber. Die Aufgabenstellung der Meta-Evaluation umfasste:

- kritische Reflexion der Vorgehensweise und Schlussfolgerungen im Hinblick auf die definierten Fragestellungen der Evaluation;
- Fokus auf Methoden/Vorgehensweise der Evaluation mit der Frage: Inwiefern waren die Methoden geeignet, um eine valide Informationsbasis zu erarbeiten und auf Basis der Ergebnisse begründete Empfehlungen auszusprechen;
- darüber hinaus Bewertung, inwieweit die vorgelegten Dokumente aktuellen Standards und Methoden der empirischen Sozialforschung entsprechen.

Zunächst machte die Referentin deutlich, dass die Einhaltung der Standards für Evaluation in der Ausschreibung nicht eingefordert wurde. Engin reflektierte dann anhand sämtlicher 25 DeGEval-Standards für Evaluation, inwieweit diese für ihr Projekt der Meta-Evaluation geeignet oder handlungsleitend gewesen wären. Diese Herangehensweise ist sehr relevant, enthalten die Standards für Evaluation selbst doch den Standard G9 „Meta-Evaluation", wo es im Erläuterungsteil heißt: „Im Rahmen einer Meta-Evaluation können die hier vorliegenden Standards eingesetzt werden […]." (DeGEval 2004: 36). Die Standardformulierung selbst legt implizit als wichtigste Informationsbasis für Meta-Evaluationen den Evaluationsbericht und begleitende Dokumente fest: „Um Meta-Evaluationen zu ermöglichen, sollen Evaluationen in geeigneter Form dokumentiert werden" (DeGEval 2004: 35). So stellte die Referentin heraus, dass insbesondere die Nützlichkeitsstandards und die Genauigkeitsstandards gut für die Meta-Evaluation nutzbar sind. Dies verwundert nicht, sind in diesen beiden Standardgruppen doch auch die meisten Vorgaben für die Evaluationsberichterstattung enthalten. Insbesondere N6 „Vollständigkeit und Klarheit der Berichterstattung" ist für Engin im Kontext der Meta-Evaluation zentral. Schließlich ist der Evaluationsbericht doch faktisch der zentrale Evaluationsgegenstand. Gerade der angesprochene Standard G9 „Meta-Evaluation" erfordere die Orientierung am Standard N6. Dasselbe gelte für die gesamt Gruppe Genauigkeit, insbesondere aber für die Standards G6 „Systematische Fehlerprüfung" und G7 „Analyse qualitativer und quantitativer Informationen". Die Referentin konstatierte überhaupt deutliche Überschneidungen zwischen den Standardgruppen N und G, die auch als Leitlinien bei der Erstellung von Evaluationsberichten dienen könnten.

Auf der anderen Seite eigneten sich verschiedene andere Standards kaum zur qualitativen Bewertung eines Evaluationsberichts – und könnten dementsprechend auch nicht eingehalten werden. So sei Standard D2 „Diplomatisches Vorgehen" kaum aus einem Bericht heraus zu beurteilen. Dasselbe gelte für Standard F2 „Schutz individueller Rechte", N3 „Glaubwürdigkeit und Kompetenz der Evaluierenden" und N5 „Transparenz von Werten". In der Anforderung von N5 „Die Perspektiven und Annahmen der Beteiligten und Betroffenen, auf denen die Evaluation und die Interpretation der Ergebnisse beruhen, sollen so beschrieben werden, dass die Grundlagen der Bewertungen klar ersichtlich sind" sieht Engin zwar eine gute Grundlage für die Meta-Evaluation, würde diesen Standard aber eher N6 „Vollständigkeit und Klarheit der Berichterstattung" zuschlagen. Ansonsten seien Perspektiven, Werthaltungen und implizite Annahmen aus einem Evaluationsbericht schwer herauszulesen.

Die Referentin zog weitere Schlussfolgerungen, die teilweise auch für den aktuellen DeGEval-Revisionsprozess der Standards von Nutzen sein könnten. So konstatierte sie, dass die definierten Einzelstandards nicht vollständig trennscharf zueinander seien und führte als Beispiele N4 „Auswahl und Umfang der Informationen" und D1 „Angemessene Verfahren" an. Eingedenk der Tatsache, dass die Standards sich gegenseitig bedingen und teilweise in einem hierarchischen Verhältnis zueinander stehen, schlug sie vor, die Standards in ein logisches Modell oder ein Zielsystem zu überführen. Als Beispiel führte sie auf: Leitziel N8 „Nutzung und Nutzen der Evaluation", Mittlerziel N6 „Vollständigkeit der Berichterstattung" und als Handlungsziele die Genauigkeitsstandards.

Die regen Diskussionen reflektierten die Möglichkeiten und Grenzen der Standards für Evaluation und zeigten Verbesserungspotenziale auf. Hierzu in der gebotenen Kürze einige Schlaglichter:

Die Standards wurden als Reflexionsinstrument für eigene Evaluationsprojekte von den Referent(inn)en durchaus gewürdigt. Ferner stand das Thema Ziele im Fokus, zum einen die Beachtung der Ziele des zu evaluierenden Gegenstands, wie sie im Standard G1 „Beschreibung des Evaluationsgegenstandes" nicht angesprochen ist. Zum anderen wurde gefordert, dass die Standards selbst mehr als Ziele formuliert werden sollten, siehe z.B.N2 „Klärung der Evaluationszwecke" und G3 „Beschreibung von Zwecken und Vorgehen".

Das Thema „Auftraggebende – Auftragnehmende" wurde verschiedentlich diskutiert: Hier bestehe ein „Ungleichgewicht", da vor allem die Auftragnehmenden angesprochen werden, obwohl die Standards auch für die Auftraggebenden gelten. Es wurde für den Revisionsprozess angeregt, den Erläuterungsteil zu ergänzen im Sinne: Was bedeutet dieser Standard für den Auftragnehmenden, was bedeutet er für den Auftraggebenden? Weitere Forderungen waren, dass Auftraggebende und Auftragnehmende im Vorfeld eines Evaluationsprojekts die Standards durchgehen sollten. Hier wurde auch ein Verfahren, um zu einem Konsens zu kommen, propagiert.

Hinsichtlich des Standards G1 „Beschreibung des Evaluationsgegenstandes" sollte – so ein(e) Teilnehmer(in) – auch dargelegt werden, was alles nicht beschrieben wird.

Die von Engin beschriebene Hierarchie innerhalb der Standards – obwohl diese eigentlich nicht intendiert ist – wurde als sinnvoll erachtet. Ein anderer Teilnehmer regte an, dass Formulierungen wie „soweit wie möglich“ ersetzt werden sollten durch „angemessen“. Schließlich wurde das funktionale Inhaltsverzeichnis als nicht sehr hilfreich bezeichnet.

Literatur

DeGEval – Gesellschaft für Evaluation (Hg.) (2004): Standards für Evaluation. Alfter (3. Aufl.).