10 Todsünden in der Evaluation interaktiver Lehr- und Lernmedien

Peter Baumgartner

Baumgartner, P. 1999. 10 Todsünden in der Evaluation interaktiver Lehr- und Lernmedien. In: Studieren 2000 - Alte Inhalte in neuen Medien? (mit CD-ROM), Hg. von K. Lehmann. Münster: Waxmann. 199-220.

1. Evaluation als BeWERTungsverfahren

Selbstverständlich ist der Titel dieses Artikels nicht wörtlich zu nehmen, sondern in mehrfacher Hinsicht eine metaphorische Übertreibung. Nicht alles was hinkt, ist ein Vergleich: Weder sehe ich mich als Papst (oder gar Gott) der Medienevaluation, noch möchte ich moralische (Reue-)Verhältnisse suggerieren. Warum also dann überhaupt der großspurige Titel?

Ich möchte in diesem Beitrag 10 Fehler in der Evaluation mediengestützten Lernens darlegen, die so grundlegend (aber auch weitverbreitet) sind, daß eine Art von Stigmatisierung durchaus sinnvoll wäre. Dabei handelt es sich nicht um "Fehler" im traditionellen Sinn, die – wenn sie uns bekannt und bewußt sind – einfach vermieden oder unterlassen werden können, sondern um systematische Verhaltensweisen, die einer inneren Logik folgen.

Die "Sünden", die ich im folgenden aufzeigen möchte, lassen sich alle auf das gleiche grundlegende mentale Modell von "Evaluation" zurückführen: auf eine Haltung, die sich aus einer dem positivistischem Wissenschaftsverständnis verpflichteten Sichtweise bzw. Definition von Evaluation ergibt.

Ich habe bereits an anderer Stelle ausführlich begründet, warum Evaluation vor allem als ein Prozeß der Beurteilung und Bewertung betrachtet werden muß (Baumgartner und Payr 1996; Baumgartner 1997 und 1999):

> Evaluation is the determination of a thing's value. (Worthen und Sanders 1987:22)
>
> Evaluation is the process of determining the merit, worth and value of things, and evaluations are the products of that process. (Scriven 1991b:1)

Aus dieser Sichtweise ergibt sich nicht nur eine trennscharfe Taxonomie von Evaluationsansätzen, sondern auch eine spezifische innere Logik (Struktur) des Ablaufes von Evaluationen:

- *Formulierung von Wertkriterien:* In der ersten Phase werden jene Kriterien ausgewählt und definiert, die der Evaluand (die evaluierte Sache, der evaluierte Prozeß etc.) erfüllen muß, um als gut, wertvoll etc. gelten zu können.
- *Formulierung von Leistungsstandards:* Für jedes einzelne Kriterium wird eine Norm definiert, die der Evaluand erreichen muß, damit das Kriterium als erfüllt angesehen werden kann (Operationalisierung).
- *Messung und Vergleich (Analyse):* Nun wird jedes Kriterium beim Evaluanden untersucht, gemessen und mit den jeweils vorgegebenen Leistungsstandards verglichen.
- *Werturteil (Synthese):* In dieser letzten und wohl schwierigsten Phase von Evaluationen müssen die verschiedenen Ergebnisse zu einem einheitlichen Werturteil integriert werden.

In diesem Beitrag möchte ich nun zeigen, was passiert, wenn diese

Grundhaltung (Evaluation ist Bewertung) nicht strikt eingehalten wird. Dementsprechend sind die nachfolgenden Gedanken nicht nur für die Evaluierung mediengestützten Lernens, sondern für alle Formen und Inhalte von Evaluationen relevant.
Entsprechend der obigen Ablauflogik lassen sich grob vier unterschiedliche Gruppen von "Sünden" unterscheiden:

- Fehler bei der Formulierung von Wertkriterien
- Fehler bei der Formulierung von Leistungsstandards
- Fehler bei der Messung und beim Vergleich (Analysefehler)
- Fehler bei der Erstellung des Werturteil (Synthesefehler)

2. "Sünden" beim Generieren von Wertmaßstäben

2.1 Es wird auf einen Wertanspruch verzichtet

Eine der grundlegendsten Evaluationsünden besteht darin, daß überhaupt auf jegliche Bewertung verzichtet wird. Ursache dafür ist meist ein kritisch-rationales bzw. positivistisches Wissenschaftsbild *und* eine implizite Gleichsetzung von Evaluations- und Wissenschaftslogik.

Meistens wird die fehlende Entwicklung von Wertmaßstäben nicht einmal besonders begründet oder diskutiert. Ausgehend von dem (falsch verstandenen) Weber'schen Postulat der Wertfreiheit (Weber 1988a,b) wird eine mögichst systematische (genaue, umfassende, relevante usw.) Erfassung von Daten versucht. Dies läuft letztlich auf eine bloß „objektive" d.h. „intersubjektive" Beschreibung eines Sachverhaltes (des Evaluanden) hinaus. Nach dem Motto „give them just the facts" wird die eigentliche Aufgabe der Evaluation, die *Bewertung* der Fakten bzw. des Datenmaterials, nicht durchgeführt.

Es lassen sich drei Erscheinungsformen dieser Wertaskese unterscheiden:

- Evaluation wird als reine Datenanalyse verstanden und mit der Konstruktion und Auswertung von Tests (bzw. anderer quantitativer Meßverfahren) gleichgesetzt.
- Evaluation wird bloß als ein Bündel von Methoden („Methodenlehre") betrachtet, die es gilt „richtig" anzuwenden.
- Evaluation wird bloß als ein (weiteres) Anwendungsgebiet der Sozialforschung gesehen und ist von daher dem Prinzip der Trennung von Beschreibung und Werturteil in den Sozialwissenschaften verpflichtet.

Manchmal wird auch eine detaillierte Datenanalyse und statistische Interpretation mit dem Aufstellen eines Wertmaßstabes und einer Wertzuweisung verwechselt. Unter Bewertung ist jedoch hier nicht bloß eine statistische Interpretation gemeint, sondern vor allem die Entscheidung darüber, ob und wie weit der Evaluand den Wünschen bzw. Vorstellungen entspricht oder aber modifziert bzw. gar gestoppt, eingestellt, aufgelassen etc. werden soll. Die Begriffe „Entscheidung" und „Wunsch" zeigen mit aller Deutlichkeit *psychologische* Gebiete auf, die weder durch Statistik noch durch eine umfassende Methodenlehre (Methodologie) abzudecken sind.

Für eine umfassende Sichtweise von Evaluationen als Bewertungen müssen fünf Gruppen von grundsätzlichen Fragen gestellet werden:

- Welche inhaltlichen Probleme deckt der Evaluand ab? Kann das durch den Evaluanden abgeckte Bedürfnis anders besser befriedigt werden?
- Wie werden gültige Fakten zur Analyse und Bewertung gewonnen?

(Erkenntnistheorie, Wissenschaftstheorie)
- Wie ist der Evaluand zu bewerten? (Werttheorie)
- Wie können die Ergebnisse der Evaluation umgesetzt werden? (Theorie über gesellschaftlichen Wandel)
- Welches pragmatische Design soll für die Evaluation gewählt werden? (Angewandte Methodologie)

2.2 Ein impliziter Wertanspruch wird nicht hinterfragt

Besonders schwer zu durchschauen ist das Fehlen von Werten dann, wenn es um die Verbesserung eines scheinbar einleuchtenden Ziels in der Evaluation geht. In diesem Falle wird Evaluation bloß als Verbesserung praktischer Maßnahmen (Treatments) betrachtet. Werden Evalutionen auf die Erfordernisse eine kurzsichtigen Praxis zurechtgestutzt und als impliziter Wertmaßstab für eine gelunge Evaluation das Generieren von Verbesserungsvorschlägen gesehen, so ist dies jedoch gleich in dreifacher Hinsicht problematisch:

Erstens können Evaluationen auch für bloße „go/stop"-Entscheidungen sinnvoll durchgeführt werden: Soll z.B. eine bestimmte Maßnahme fortgeführt oder abgebrochen werden?

Zweitens aber impiziert die Entwicklung von Verbesserungsvorschlägen eine ganz andere Logik als sie Evaluationen im allgemeinen innewohnen: Im Prinzip geht es bei Evaluationen um die Erstellung und Zuweisung eines Wert*urteils* (Evaluand = gut/ schlecht, wertvoll/wertlos). Bei der detaillierten Analyse von Mängel handelt es sich jedoch um einen anderen *fachlichen* Inhalt. Es werden dabei andere Fachgebiete und andere Kenntnisse angesprochen: *Der Inhaltsexperte ist nicht automatisch der Evaluationsexperte und umgekehrt.* So kann z.B. eine vergleichende Produktevaluation von Computermonitoren zu klaren Ergebnissen kommen und Mängel bestimmter Markenprodukte eindeutig feststellen (z.B. zu hohe Strahlungsintensität). Wie und ob diese Mängel jedoch bei diesem Produkt behoben werden können oder sollen, ist eine ganz andere Sache und verlangt vielleicht eine weitere Studie und/oder Laborexperimente mit ganz anderen inhaltlichen Kompetenzen, als sie von EvaluationsexpertInnen benötigt werden.

Drittens aber gibt es keine scheinbar „objektiven", nicht hinterfragbaren Ziele, die quasi automatisch für sich sprechen. Auch wenn mir in Dresden ein Evaluationsexperte triumphal Praxisferne attestiert und mir als Beispiel ein klares, nicht weiter zu diskutierendes Ziel der Automobilindustrie entgegenschleudert (Verkürzung der Bremswege: „Ein kurzer Bremsweg ist einfach gut, darüber muß nicht diskutiert werden!"), so gilt trotzdem: Ziele existieren nicht im luftleeren Raum, sondern sind immer in einer Wertehierarchie eingebettet.

Tatsächlich impliziert das Ziel eines möglichst kurzen Bremsweges bereits eine ganze Reihe von (positiven) Wertvorstellungen, wie z.B. eine motorisierte Gesellschaft mit schweren, hochgezüchteten Personenkraftfahrzeugen, die aber trotzdem möglichst ohne Personen- und Sachschaden benutzt werden können sollen. Ginge es wirklich nur um die Verkürzung des *individuellen* Bremsweges eines PKW's so wären z.B. Greifmechanismen möglich, die zwar den Straßenbelag zerstören, aber das Auto dafür sehr schnell zum Stehen bringen. Wenn es um die *gesellschaftliche* Verkürzung von Bremswegen geht, wären sowohl städtebauliche Maßnahmen oder aber Verkehrsvermeidung (d.h. weniger Autos und damit gefahrene Kilometer) eine geeignete Strategie.

Wenn sich Evaluationen nicht mit Wertefragen explizit auseinandersetzen, kommen sie in Gefahr als Pseudo-Evaluationen bloß

der Erfüllung machtpolitischer Interessen dienlich zu sein.

2.3 Sünden bei der Formulierung der Wertkriterien

Je nachdem, wie Evaluationen mit der grundsätzlichen Frage der Zuweisung von Werturteilen umgehen, lassen sich bestimmte Studien als unechte Pseudo- und Quasi-Evaluationen stigmatisieren (Stufflebeam und Shinkfield 1985:45-57). In dieser Hinsicht stellt die Gleichung (Evaluation = Bewertung) bereits ein sehr scharfes Trennkriterium dar.

2.3.1 Pseudo-Evaluationen

Darunter sind alle Untersuchungen einzuordnen, die entweder politisch gesteuert sind oder ganz klar die Festigung (Bestätigung) einer vorgefaßten Meinung intendieren. Besonderes Kennzeichen dieser Art von Studien ist es, daß keine vollständige, umfassende und ausgewogene Analyse und Bewertung vorgenommen wird. Ausgangspunkt dieser Erhebungen sind:

- Die möglicherweise bei einer echten Evaluation gefährdete Position einer Adressatengruppe führt zu einem Interessenskonflikt. Eine Pseudoevaluation soll daher Argumente für diese Unischerheit liefern und so die damit verbundene Interessensgruppierung stärken (= *politisch kontrollierte Studie*). Meistens werden diese Art von Untersuchungen verdeckt durchgeführt. Dadurch wird einerseits vermieden, daß die Öffentlichkeit vorzeitig ihre Aufmerksamkeit auf die unter Druck geratene Position lenkt. Andererseits bleibt die Studie – falls ihre Ergebnisse den Auftragebern nicht entsprechen – in der Schublade und wird nicht veröffentlicht.
- Der Versuch, durch gezielte Verbreitung bestimmter Informationen andere Interessensgruppierungen in ihrem Verhalten zu beeinflussen. Meistens dient sie dazu, ein bestimmtes Objekt (z.B. Konsumprodukt) in einem besonders vorteilhaftes Licht erscheinen zu lassen (= *Public-Relation Studie*). Besondere Kennzeichen dieser Studien sind ihre methodologische Fragwürdigkeit („quick and dirty"), die meistens zu einem (intendierten) systematischen Fehler führen.

Pseudo-Evaluationen geben nur vor, Evaluationen zu sein. Sie versuchen die Autorität von echten Evaluationen für ihre eigenen (dubiosen) Interessen einzusetzen.

2.3.2 Quasi-Evaluationen

Hierbei handelt es sich um Untersuchungen, die zwar methodisch korrekt durchgeführt werden, jedoch bereits eine eingeschränkte – nicht mehr weiter zu hinterfragende – Ausgangsfragestellung haben. Besonderes Kennzeichen dieser Analysen ist es, daß sie eine Begründung, Diskussion und eventuelle Kritik der aufgestellten Wertansprüche vernachlässigen oder aber kritiklos zulassen. Sie nehmen die Aufgabenstellung unhinterfragt hin und beschäftigen sich sogleich mit der Auswahl einer adäquaten Methode zur Untersuchung der Problematik. Typische Beispiele für Quasi-Evaluationen sind:

- *Ziel-orientierte Evaluationsansätze* wie sie z.B. von Ralph Tyler in den 30-er Jahren entwickelt worden sind. Dazu ist auch die von Provus entwikkelte Diskrepanz-Analyse zu zählen. Ausgehend von breit formulierten Zielen, die dann verfeinert und operationalisierbar gemacht werden, sollen Diskrepanzen zwischen Ziel und Realisierung festgestellt werden. Im Extremfall – wie z.B. bei gewissen Management-Informationssystemen (MIS) wird nur mehr beobachtet, ob der Evaluand gewisse Minimalkriterien überschreitet bzw. erfüllt (*monitoring*). Obwohl zielbasierte Ansätze scheinbar objektiv sind,

bedeuten sie immer eine Art von Tunnelvision, weil nur mehr vorgegebene Ziele untersucht werden. Damit ist die Legitimität der Untersuchung gefährdet, außerdem bleiben nicht intendierte Effekte unberücksichtigt. Scriven (1991a) schlägt daher vor, diese mögliche Verzerrung (*bias*) durch eine ergänzende zielfreie Evaluation (*goal-free evaluation*) zu korrigieren. Dabei wird der Evaluand völlig unvoreingenommen untersucht. Offizielle Ziele, programmatische Papiere, Meinungen des Staffs und des Management etc. werden in dieser ersten Phase absichtlich *nicht* erhoben.

- Experimentelle Untersuchungen wie sie z.B. im quasi-experimentellen Forschungsdesign (Vergleichsgruppen) üblich sind (vgl. Thorndike et al. 1991; Wiersma 1991). So wird beispielsweise der Lernerfolg zweier Gruppen untersucht, die unterschiedlichen Maßnahmen (*treatments*) ausgesetzt worden sind (z.B. traditioneller Unterricht versus Verwendung von interaktiver Software). Die sorgfältige Beachtung methodologischer Forderungen (Vortest, Ähnlichkeit der beiden Gruppen in anderen als der untersuchten Variablen wie z.B. Alter, Geschlecht etc.) verhindert nicht, sondern begünstigt die Mißachtung der ihnen implizit zugrunde liegenden Ziele. Werturteile werden unwidersprochen und z.T. sogar unbewußt akzeptiert. Was gilt z.B. als Kriterium für einen Lernerfolg? Ist es wirklich die bloße Erinnerung bei einem multiple- choice Test oder die richtige und vollständige verbale Reproduktion der vermittelten Inhalte bei offenen Fragen? Obwohl die komplexen Untersuchungsinstrumente (wie z.B. Fragebogen) zwar methodisch einwandfrei konstruiert worden sind, messen sie immer nur das, was bereits als Ausgangspunkt ihrer Konstruktion unhinterfragt angenommen wurde („methodischer Zirkelschluß“). Und das kann oft auch völliger Unsinn sein („garbage in - garbage out“)

Im Gegensatz zu den Pseudo-Evaluationen können Quasi-Evaluationen durchaus ihre Berechtigung haben und im Einzelfall sogar sehr wertvoll sein. Sie klammern jedoch sowohl grundsätzliche Fragen zu den Zielsetzungen und den damit verbundenen Werturteilen als auch moralische Aspekte aus und sind oft interessensdominiert.

2.3.3 Akzeptanzstudien

Ein anderer Fehler bei der Formulierung von Wertkritieren besteht darin, daß bloß die augenblicklichen und aktuell gültigen Wertmaßstäbe erhoben werden. Statt sich seitens der Evaluatoren *vor* dem Evaluationsverfahren zu überlegen, welche Kriterien einen Evaluanden als gut auszeichnen, und *danach* erheben, ob, wo und inwieweit diese Kriterien auch tatsächlich erfüllt werden, wird bloß das Publikum (z.B. Kunden, Adressaten etc.) befragt. Statt ein begründetes Werturteil aufzustellen und dessen Realisierung zu untersuchen, wird bloß erhoben, inwieweit vorhandene (Wert-)Ansprüche befriedigt werden - unabhängig davon, ob sie legitim (begründbar) sind oder nicht. Statt einer echten Evaluation, wird bloß eine Zufriedenheitsstudie durchgeführt.

Selbstverständlich haben gerade auch in der Lehre solche Untersuchungen ihre Berechtigung. Schließlich ist die Erhebung der subjektiven Urteile der Klientel (z.B. Studierende) ein ganz wichtiger Faktor der Bewertung der Lehre. Allerdings ist vor einem vollständigen Aufgehen in eine marktorientiere Service- bzw. Kundenorientierung zu warnen: Erstens sind eher die Betriebe (und nicht die Studierenden) die eigentlichen Kunden der Wissensproduktion an den Hochschulen (indem sie die Absolventen einstellen oder eben nicht einstellen), zweitens funktionieren die gängigen Marktmechanismen beim Erwerb von Wissen und Fertigkeiten bzw. beim Erkenntnisakt nur sehr bedingt.

Jede noch so gute Serviceleistung (z.B. Verteilung von Skripten) scheitert, wenn sie nicht von einem Akt der indviduellen persönlichen Erkenntnis (Polanyi 1962, 1969) genützt wird. Auch wenn wir durch unsere Reden, Schriften oder andere Medien auf etwas hindeuten können (deiktische Definition), so bleibt die Erfassung der Gestalt, die Anwendung, Einverleibung, „the knack of it" der Anstrengung des einzelnen Indviduums, dem Studierenden vorbehalten. Und dieser Akt des Verstehens ist gerade *nicht* umgekehrt proportional der ihm vorgelagerten deiktischen Anstrengungen (vgl. dazu Baumgartner 1993).

2.4 Sünden bei der Zuweisung von Werten

2.4.1 Kategorienfehler

Ein Kategorienfehler wird dann begangen, wenn Begriffe auf unnterschiedlichen Ebenen miteinander verglichen bzw. in Beziehung gesetzt werden. Gilbert Ryle (1969) bringt dafür anschauliches Beispiel aus dem Alltag: Einem Besucher werden alle Gebäude der Universität (Hörsäle, Rektorat, Dekanat Mensa, Studienberatung,…) gezeigt. Nachdem der Besucher alles eingehend betrachtet, besucht und studiert hat, fragt er uns: „Schön ich habe jetzt viele Gebäude und Räumlichkeiten gessehen, aber wo ist die Universität?"

Diese Frage ist nicht zulässig, bzw. macht keinen Sinn weil sie zwei grundversschiedene Ebenen in Beziehung zueinander setzt: Die „Universität" als abstraktes Gebilde mit ihren Prüfungs- und Studienordnungen, mit ihren sozialen Settings und Rollen (Professor, Studierende) läßt sich nicht auf der räumlichen Ebene erfassen.

Genauso wie bei diesem (trivialen) Beispiel verhält es sich jedoch auch beim Verhältnis von prodzeduralem Wissen („Wissen, wie" oder *know how*) und Fertikeiten (*abilities*, *skills*). Das „Wissen, wie" ist immmer noch (ähnlich, wie das „Wissen, daß" etwas der Fall ist) grundsätzlich ein theoretisches Wissen und keine praktische Fertigkeit. Wie beim Universitätsbeispiel machen daher bestimmmte Fragen keinen Sinn. So sind z.B. Handlungen nicht äquivalent in Worten faßbar – auch wenn dies Habermas in seiner Theorie des kommunikaten Handelns behauptet (1981 und 1984. Vgl. zur ausführlichen Kritik dazu Baumgartner 1993:152ff.). So antwortet die berühhmte tänzerin Isodora Duncan auf die Frage, was ihre Tänze zu bedeuten hätten: „If I could tell you what it meant, there would be no point in dancing it" (zitiert nach Bateson 1972:137 und 464).

Der Kategorienfehler hat weitreichende – unter anderem auch (prüfungs)didaktische – Konsequenzen: Ist das, was gemessen wird auf derselben Ebene wie das was eigentlich beurteilt werden soll? Meistens wird – damit das Ziel der intersubjektiven Überprüfbarkeit erreicht wird – durch genau definierte Zuschreibung gemessen. Wie methodisch genau auch diese Zuschreibung erfolgen mag, sie ist und bleibt eine Zuschreibung von außen, eine Zuschreibung der dritten Person, die zu den im inneren einer Person staffindenen (Lern-)Vorgängen eine andere Qualität (Ebene) darstellt.

Georg Neuweg geht diesem Kategorienfehler in all seinen Verästelungen und Konsequenzen in seiner (noch nicht) erschienenen Habilationsschrift nach (1998). Es wäre eine äußerst lohnenswerte Aufgabe seine im Zuge der Rezeption von Michael Polanyi vorwiegend erkenntnis- und wissenschaftstheoretischen Äußerungen auf praktische Konsequenzen im Evaluationsbereich anzuwenden. Das muß hier sowohl aus Platz- und Zeitgründen unterbleiben. Festzuhalten aber ist:

Die Verwechslung von Denkprozessen und Denkprodukten ist ein Kategorienfehler und stellt einer der schwerwiegensten Fehler bei

Evaluationen zum Lernerfolg dar.

2.4.2 Skalenfehler

Für die Zuweisung von Werten (Beurteilungsverfahren) lassen sich grundsätzlich vier Methoden unterscheiden:

a) Einstufung (grading):
Die Beurteilung findet an Hand eines vorweg definierten Bewertungsmaßstabes statt. Dies ist z.B. dann der Fall, wenn bei einer Klausur die Beurteilung streng nach der Anzahl der beantworteten Fragen1 erfolgt. Häufig wird jedoch von den Lehrkräften gegen diesen Verfahren aus optischen Gründen gesündigt. Die Verteilung wird meist so „nachgebessert", daß etwa eine Normalverteilung entsteht („grading on the curve"). In diesem Fall handelt es sich jedoch nicht mehr um Einstufung, sondern um Reihung.

b) Reihung (ranking):
Hier werden die Evaluanden relativ zu einander beurteilt. Es entsteht eine Reihefolge (gut-besser-am Besten, häufig-selten-nie usw.). Weder zu den Abständen der Evaluanden untereinander noch zum Ausmaß der Werterfüllung kann eine gesicherte Aussage gemacht werden (Ordinalskala).

c) Punktevergabe (scoring):
Wenn Punkte vergeben werden, so ist unbedingt darauf zu achten, daß die Abstände zwischen den einzelnen Punkten bedeutungsvoll und äquidistant sind (Intervall- oder falls es einen Nullpunkt gibt Ratio-Skala, auch metrische Skala genannt), ansonsten handelt es sich um bloßes Ranking. Unsere Schulnoten z.B. stellen bloß eine Reihung dar: Operationen wie Addition, Division (wie sie z.B. für den Notendurchschnitt berechnet werden) sind – streng gesehen – nicht zulässig.

d) Aufteilung, Zuteilung (apportioning):
Hierbei werden vorhandende Ressourcen entsprechend der Wertigkeit der Evaluanden aufgeteilt (z.B. Zuteilung von Budgetmitteln). Es ist eine häufige Praxis begrenzte Ressourcen durch ein scheinbares Ranking zu verstecken und somit nur den oberen Plätzen eine Leistung zuzuteilen.

Selbstverständlich können die verschiedenen Methoden der Wertzuweisung auch kombiniert vorkommen: So könnte theoretisch z.B. bei Sportereignissen alle Bewertungsverfahren von der Einstufung (Ausscheidung, Qualifikation) über Reihung oder Punktevergabe (z.B. bei Zeit-, Gewicht- oder Längenmessungen) bis zur Zuteilung (Preisverleihung) angewendet werden.

1Wir wollen hier der Einfachheit halber annehmen, daß alle Fragen gleiche Wertigkeit haben, dh. gleich gewichtet sind.

2.5 Sünden bei der Gewichtung von Wertansprüchen

Es ist inzwischen deutlich geworden, daß für eine ordentliche Analyse des (meistens äußerst komplexen) Evaluanden verschiedene Faktoren bzw. Komponenten betrachtet werden müssen. Ein wichtiges Problem hierbei ist die Festlegung der relativen Wertigkeit (Gewichtung) dieser verschiedenen Dimensionen.

Wenn wir vorerst den inhaltlichen Zusammenhang zwischen Funktionsmerkmalen des Evaluanden und Intererssensorientierungen verschiedener Adressaten der Evaluation ausklammern, so stellt sich das Definieren von Prioritäten (Gewichtungsproblem) zuerst einmal als methodisches Problem dar. Im Prinzip gibt es zwei Verfahren: additive (numerische) und qualitative Gewichtungsprozeduren.

2.5.1 Numerische Gewichtung und Summierung (NGS)

Es stellt derzeit das dominante Modell für eine komplexe Produktevaluation dar und wird insbesondere im Zusammenhang mit der Bewertung von Lernsoftware in Form von Check- oder Prüflisten angewendet (vgl. Baumgartner 1995; Biermann 1994; Doll 1987; Fricke 1995; Thomé 1988). Numerisches Gewichten und Summieren (NGS) kommt in verschiedenen Formen vor und kann sowohl beschreibend, vorschreibend (normativ, präskriptiv) oder auch bewertend eingesetzt werden. Die allgemeine Form ist die *Multi-Attribute Utility Analysis* (Scriven 1991a:380f.):

- Zuerst werden die einzelnen Dimensionen in ihrer relativen Wertigkeit (z.B. anhand einer 1-3-, 1-5 oder 1-10-Skala) eingeschätzt (gewichtet).
- Anschließend wird die Leistung des Evaluanden nach den einzelnen Dimensionen eingeschätzt (rating).
- Das Produkt von Leistungsbewertung und Gewicht (Leistungspunkte x Gewichtung) wird berechnet und für jeden einzelnen Evaluanden summiert.
- Es ergibt sich für jeden Evaluanden eine einzige Zahl, die den relativen Rang des jeweiligen Evaluanden bestimmt. Sieger ist der Evaluand mit der größten Punktezahl.

Das NGS-Verfahren ist infolge einer Reihe von Vorteilen (leicht verständlich, einfach durchzuführen, immer aufschlußreich, ergibt manchmal auch valide Ergebnisse) sehr beliebt. Obwohl es immer einen ersten Aufschluß bzw. Einblick bietet und daher im Rahmen einer weiterführenden Evaluation durchaus brauchbar ist, hat es schwerwiegende intrinsische methodische Mängel, sodaß der alleinige Rekurs auf dieses Verfahren verboten werden sollte:

- Ein Set von Gewichten löst nicht das Problem, daß einige Dimensionen (Merkmale) erst dann eine sinnvolle Funktion des Evaluanden darstellen, wenn ein bestimmtes Mindestmaß überschritten ist. In einer abschließenden Summierung zu einer einzigen Zahl gehen diese inhaltlichen Minimalanforderungen jedoch verloren. Diese Schwierigkeit läßt sich jedoch durch eine Erweiterung des NGS-Verfahrens beheben (NGS-Modell mit Minima)2.
- Ein weiteres (lösbares) Problem besteht darin, daß die einzelnen Bewertungskomponenten der Evaluanden oft nicht unabhängig voneinander zu betrachten sind, weil sie miteinander interagieren.

2Dazu wird jeder der Kriterien, das ein bestimmtes Minimum erfüllen muß, zuerst geprüft, bevor weiter analysiert wird. Nur die Leistung über dem Minimum wird danach gewichtet. Evaluanden, bei denen einzelne Kriterien dieses notwendige Minimum nicht erreichen, scheiden aus.

Diese Schwierigkeit könnte durch eine Neubestimmung bzw. neue Definition der Kriterien gelöst werden. Allerdings ist dies nicht immer einfach, erfordert große Geschicklichkeit und Kenntnisse und stellt fast immer nur eine ad hoc-Lösung dar, die nicht verallgemeinert werden kann.

- Eine wesentliche Kritik an der NGS-Methode besteht darin, daß sie eine lineare Skala der Nützlichkeit (Vergabe von Punkten und Summierung) annimmt, was jedoch sicherlich falsch ist! Die verschiedenen Komponenten des Evaluanden lassen sich nicht über eine einzige Skala bewerten. So macht es z.B. wenig Sinn Kriterien der Benutzeroberfläche und Interaktivität von Lernsoftware in einer einheitlichen Skala zu summieren. Das wäre nur dann sinnvoll, wenn diese Kriterien für den Lernerfolg die gleichen Auswirkungen hätten, also auf einer linearen Skala liegen würden. Das wurde aber in keinem einzigen Fall bisher theoretisch nachgewiesen!
- Ähnlich wie beim erweiterten NGS-Modell (mit Minima) läßt sich auch beim Problem der linearen Skala der Nützlichkeit durch ein sequentiell durchgeführtes Ausscheidungsverfahren provisorisch „Nachbessern": Es wird die Liste der Merkmale nicht zufällig (alphabetisch oder nach einer anderen inhaltlich irrelevanten Reihenfolge) durchgearbeitet, sondern zuerst werden die absoluten Notwendigkeiten festgestellt und dann so viele Kandidaten wie möglich eliminiert. Allerdings bleibt die grundsätzlich falsche Annahme einer Linearität der Punkteabstände im weiteren Verfahren bestehen. Multiplikation und Summenbildung sind nur bei Intervall- oder Ratio-Skalen zulässige Operationen, während es sich hier um eine Ordinalskala handelt, die nur eine Reihung der einzelnen Merkmale erlauben würde.
- Die entscheidende Kritik bzw. das (unlösbare) Hauptproblem des NGS-Verfahren besteht jedoch darin, daß die Anzahl der Kriterien nicht voraussehbar ist. Sie kann von etwa einem Dutzend bis zu einigen hundert Kriterien reichen. Damit werden aber entweder wichtige Dimensionen durch eine Vielzahl von Trivialitäten überschwemmt oder aber weniger wichtige Faktoren wirken sich auf das Gesamtergebnis zu stark aus. Das Festlegen einer fixen Punkteanzahl, die nicht überschritten werden darf, reduziert zwar das Problem, kann es aber nicht gänzlich lösen. Was sind die relevanten Kriterien (wie viele, wie detailliert) und welche Gewichtung kommt ihnen jeweils zu?

Besonders fatal beim NGS-Verfahren ist es, daß diese Gewichtungsprozedur keine Spuren hinterläßt. Da sich als Ergebnis bloß eine *einzige* Zahl pro Evaluand ergibt, sind nachträglich keine inhaltlichen Fehlerkorrekturen mehr möglich.

2.5.2 Qualitative Gewichtung und Summierung (QGS)

Obwohl mit dem NGS-Verfahren zwar viele Evaluanden in einem ersten Durchgang provisorisch miteinander verglichen werden können und es als erster grober Filter durchaus brauchbar ist, ist es letztlich doch notwendig, einen *paarweisen Vergleich mit qualitativen Bewertungsverfahren* durchzuführen (Scriven 1991a:293ff.):

- In einem ersten Schritt werden für die einzelnen Dimensionen nur fünf Gewichte vergeben. Es empfiehlt sich dafür Symbole zu verwenden, damit gleich von vornherein eine Verwendung als Intervall- oder Ratioskala ausgeschlossen wird. Bewährt hat sich folgende Einteilung: Essential (E)/ Very Valuable (*)/Valuable (#)/Marginally Valuable (+)/Zero (0).Damit wurde nicht nur die Gewichtung der einzelnen Merkmale festgelegt, sondern auch festgelegt, welche Eigenschaften Minimalerfordernisse darstellen (*Essentials*).

- The rationale for this approach is that validity in allocating utility points is hard to justify beyond this very modest level – in fact, some research suggests that even a single category may be enough. But if one feels differently, one can allocate an accent (represented by the single quote, ', to indicate 'something more' than the utility symbol to which it is attached, giving six operating levels after the E and 0 filters are applied. (ebd., 294)

- Alle 0-Dimensionen können nun gestrichen werden. Sind sind als völlig unbedeutend gewichtet worden und daher für die Bewertung irrelevant. Damit wird unnötiger Analyseaufwand vermieden.
- Es wird nun überprüft, ob alle Evaluanden die Minimalerfordernisse (Kriterien, die mit E gewichtet wurden) auch tatsächlich erfüllen. Falls nicht, werden sie aus der weiteren Analyse ausgeschieden. Dadurch wird der weitere Arbeitsaufwand beträchtlich reduziert. Allerdings ist dafür Sorge zu tragen, daß es sich dabei um ein diskretes (alles-oder-nichts-Attribut) handelt (Software ist z.B. lauffähig oder nicht). Andernfalls muß das (Anspruchs-)niveau, das unbedingt erforderlich ist, genau festgelegt werden und geprüft werden, ob der betreffende Evaluand dieses Anspruchsniveau erreicht oder nicht.3
- Die verbleibenden Evaluanden weisen jetzt nurmehr Unterschiede zwischen * und + auf und werden nun im Rahmen von 0 bis zur maximalen Gewichtung des jeweiligen Kriteriums bewertet. D.h. ein #-Kriterium kann keinen höheren Wert als # erhalten (also nur 0,+,#). Es besteht jedoch keine unbedingte Notwendigkeit, Bereiche für jeden Nützlichkeitslevel zu spezifizieren, einige können auch übersprungen werden So ist es z.B. möglich, daß eine Dimension nur + und * kennt, der Bereich mit # wird überprungen. Zu beachten ist auch, daß es Fälle gibt, wo es keinen monotonen Zweckmäßigkeitsbereich gibt, d.h. wo das Überschreiten eines bestimmten Niveaus wiederum zu einer Schwäche wird (z.B. das geringe Gewicht eines Telefons, wenn es beim schnellen Abheben des Hörers vom Tisch kippt). Falls es Unsicherheiten über Zuverlässigkeit (Reliabilität) der Einschätzung eines Leistungsmerkmales gibt, kann das Symbol eingeklammert werden. Damit kann der Evaluator/die Evaluatorin die Sicherheit der jeweiligen Beurteilung ausdrücken und das betreffende Kritierium wird damit für eine spätere – eventuell notwendig gewordene – genauere Untersuchung markiert.
- Nach den bisherigen Verfahrensschritten entsteht nun eine Rangordnung (*ranking*), die anschließend auch mit einer integrierenden Schlußbewertung (*grading*) versehen werden kann (ist zu kaufen, kommt ins Finale etc.). Diese ließe sich z.B. durch das Festlegen einer Minimumanzahl von * oder * und # oder auf einer *individuellen Fallbasis* durchführen, nachdem alle Evaluanden bereits bewertet wurden. In einem disjunktiven Modell könnte auch argumentiert werden, daß alle Merkmale, die über einem gewissen Minimum liegen, die Anforderungshürde überwunden haben. Das würde jedoch die Anwendung eines *cutoff-Kriteriums* sowohl für die Gesamtbewertung als auch für jede einzelne Dimension bedeuten.
- Nun werden die Ergebnisse der Leistungsbewertung integriert, indem jede Kategorie mit der gleichen Wertigkeit summiert wird, d.h. man erhält drei Gesamtwerte für jeden Evaluanden (= Summe der *, Summe der # und Summe der +, mit oder ohne Akzent, mit und ohne Klammer)
- Nun werden jene Eigenschaften, die alle Evaluanden gleichermaßen

3 So kann z.B. eine Eigenschaft z.B. dieses Minimum nicht nur erreichen (=E), sondern darüber überschreiten und dann z.B. mit einem + versehen werden.

aufweisen (z.B. wenn alle Evaluanden ein bestimmtes Kritierium mit + erfüllt haben), ausgeschlossen. Damit wird der weitere Vergleich auf einer Fall-zu-Fall Basis vereinfacht.

- Es kann nun geprüft werden, ob bereits eine eindeutige Rangordnung möglich ist. Eindeutig heißt, daß z.B. ein Evaluand mit 3*, 4# und 2+ auf jeden Fall besser ist als einer mit 2*, 5# und 2+. Hat jedoch der zweite Evaluand z.B. 2*,7#, so ist keine eindeutige Entscheidung möglich und die beiden Kandidaten müssen noch genauer untersucht werden (paarweiser Vergleich).

  > You can be sure that if you get a winner on this restrictive basis (as is common), it is the winner... (ebd., 295)

Wenn nun nicht bereits entscheidbare Verhältnisse vorliegen, so kann eine neuerliche Gewichtung im Lichte der vergleichenden Bewertung konkreter Einzelfälle hilfreich sein. Neben seiner relativen Komplexität hat das QGS-Verfahren den Nachteil, daß es keinen definitiven Entscheidungsalgorithmus hat. Manchmal muß es als iterative Prozedur mehrfach durchlaufen werden und müssen im Lichte der bisherigen Analyse die Bewertungen nochmals durchgeführt werden. Das Verfahren wechselt damit ständig zwischen holistischer und analytischer Betrachtungsweise und ergibt immer sinnvolle vor allem jedoch nachvollzieh- und überprüfbare Ergebnisse.

Eine Beispiel für die praktische Umsetzung des QGS-Verfahren findet sich im Rahmen des alle zwei Jahre stattfindenden European Academic Software Awards (EASA). Eine detaillierte Beschreibung einer erfolgreichen Anwendung findet sich bei Baumgartner und Payr (1997), die das EASA-Finale 1996 in Klagenfurt beschreiben und kommentieren.

2.6 Sünden beim abschließenden Werturteil (Synthese)

Zum Abschluß möchte ich noch auf Fehler eingehen, die sich bei der abschließenden Beurteilung einer Evaluation zwangsläufig ergeben, wenn sie nicht als Bewertung konzipiert werden.

> Evaluation should not only be true; it should also be just… justice provide an important standard by which evaluation should be judged. (House 1980:121, hier zitiert nach Shadish, Cook und Leviton 1991:51)

Mit diesem Zitat zeigt sich eine weitere Anforderung, die Evaluationen zu erfüllen haben. Deutlich wird dieser erhöhte Anspruch, wenn die Forderungen eines US-Kommittees herangezogen werden, die inzwischen zu allgemein akzeptierte "Standards for Evaluation of Educational Programs, Projects, and Materials" geführt haben (zitiert nach Stufflebeam und Shinkfield 1985:9-15) Danach müssen Evaluationen 4 Kriterien(gruppen) genügen:

2.6.1 Evaluatiuonen sollen nützlich sein

Das Kriterium der Nützlichkeit soll durch die Einhaltung folgender Forderungen erfüllt werden:

- Evaluationen sollen sich an jene Personen(gruppen) richten, die entweder involviert, betroffen oder verantwortlich für die Umsetzung der Ergebnisse sind.
- Evaluationen sollen diesen Zielgruppen helfen, Stärken und Schwächen des Evaluanden wahrzunehmen.
- Die wichtigsten Ergebnisse, Fragen, Entscheidungsvorschläge sollen deutlich herausgehoben werden.
- Evaluationen sollen im allgemeinen nicht nur Feedback über

Stärken und Schwächen mitteilen, sondern auch Vorschläge zur Verbesserung beeinhalten.

Um die dazugehörigen Evaluationssünden zu vermeiden, haben sich folgende Fragestellungen als nützlich erwiesen:

- Sind die Adressatengruppen ausreichend und trennscharf identifiziert?
- Sind die Evaluatoren vertrauenswürdig und kompetent?
- Sind die Informationen in Umfang und Auswahl so aufbereitet, daß sie die wichtigsten Probleme und Interessen der Adressatengruppen ansprechen?
- Sind die Grundlagen der Evaluation (Design, Methodik, Auswertungs- und Interpretationsverfahren) dargestellt, sodaß eine ausreichende Basis für das Werturteil vorhanden ist?
- Sind die Ergebnisse der Evaluation verständlich und klar beschrieben?
- Sind die Ergebnisse in geeigneter Form an die Adressaten übermittelt worden?
- Sind die Ergebnisse so zeitgerecht, daß sie Verwendung finden können?
- Ist die Evaluation so geplant und durchgeführt worden, daß sie die Adressatengruppen zu énderungen motiviert?

2.6.2 Evaluationen sollen durchführbar sein

Das Kriterium der Durchführbarkeit soll durch die Einhaltung folgender Forderungen erfüllt werden:

- Evaluationen sollen Prozeduren anwenden, die ohne große (Um)brüche implementiert werden können.
- Evaluationen sollen so einfach und vorsichtig (diplomatisch) gestaltet werden, daß ihre Durchführung realistisch ist.
- Evalutionen sollen effizient durchgeführt werden.
- Evaluationen sollen die unterschiedlichen Interessensorientierungen beachten bzw. miteinbeziehen, damit Widerstände überwunden werden können.

Um die dazugehörigen Evaluationssünden zu vermeiden, haben sich folgende Fragestellungen als nützlich erwiesen:

- Sind die angwendeten Verfahren praktisch und daher einfach, ohne große Umbrüche durchführbar?
- Ist die Evaluation so geplant, daß ihre Durchführung realistisch ist und sie auch interessenspolitisch überleben kann, viabel ist?
- Hat sich die Evaluation ausgezahlt, dh. übersteigen die Vorteile ihrer Ergebnisse die Kosten ihrer Durchführung?

2.6.3 Evaluationen müssen gerecht (fair) sein

Das Kriterium der Gerechtigkeit soll durch die Einhaltung folgender Forderungen erfüllt werden:

- Evaluationen sollen auf expliziten (schriftlichen) Vereinbarungen beruhen, damit die notwendige Kooperation sichergestellt wird.
- Evaluationen müssen die Rechte aller betroffenen Gruppen wahren.
- Evaluationen müssen sicherstellen, daß ihre Ergebnisse ohne Zugeständnisse vorgelegt werden können.
- Evaluationen sollen sowohl Stärken als auch Schwächen des Evaluanden darlegen.

Um die dazugehörigen Evaluationssünden zu vermeiden, haben sich folgende Fragestellungen als nützlich erwiesen:

- Gibt es schriftliche Vereinbarungen Gruppierungen?
- Wird mit Interessenkonfliken offen und ehrlich umgegangen?
- Ist der Bericht offen, direkt und ehrlich auch zu den Limitationen seiner Ergebnisse?
- Werden von den betroffenen Gruppierungen das Informationsrecht der

ôffentlichkeit (unter Einschluß eventueller persönlicher Datenschutzbestimmungen) akzeptiert und sichergestellt?
- Sind alle Rechte und Datenschutzbestimmungen berücksichtigt und eingehalten?
- Werden menschliche Interaktionen während der Evaluation entsprechend gewürdigt und einbezogen?
- Ist der Bericht ausgewogen, sodaß er alle Stärken und Schwächen enthält?
- Ist die finanzielle Rechenschaftslegung sparsam und ethisch vertretbar?

2.6.3 Evaluationen sollen intersubjektiv überprüfbar sein

Das Kriterium der intersubjekktiven Überprüfbarkeit soll durch die Einhaltung folgender Forderungen erfüllt werden:
- Der Evaluand soll in seiner Entwicklung und in seinem Kontext klar beschrieben werden.
- Stärken und Schwächen des Evaluationsdesign, der Methoden und un Ergebnisse sollen klar aufgezeigt werden.
- Evaluationen sollen systematische Fehler vermeiden bzw. in Grenzen halten und diese mögliche Fehlerbandbreite aufzeigen.
- Evaluationen sollen zu gültigen und replizierbaren Ergebnissen führen.
- Um die dazugehörigen Evaluationssünden zu vermeiden, haben sich folgende Fragestellungen als nützlich erwiesen:
- Ist der Evaluand in seiner Funktion und Wirkungsweise soweit analysiert, daß über ihn ein klares Verständnis vorhanden ist?
- Ist das Umfeld, der Kontext des Evaluanden soweit analysiert, daß ein klares Verständnis über mögliche Einflüsse vorhanden ist?
- Sind die Quellen der Daten und Informatioen so ausreichend beschrieben, daß sie adäquat beurteilt werden können?
- Sind die Instrumente zur Informationssammlung so gewählt bzw. konsturiert worden, daß sie zu gültigen (validen) Daten führen?
- Sind die Instrumente zur Informationssammlung so gewählt bzw. konsturiert worden, daß sie zu zuverlässigen (reliaben) Daten führen?
- Ist die Datensammlung, ihre Verarbeitung und Auswertung so kontrolliert worden, daß die Fehlerwahrscheinlichkeit sehr gering ist?
- Ist die quantitative Auswertung der Daten systematisch und methodisch korrekt durchgeführt worden?
- Ist die qualitative Auswertung der Daten systematisch und methodisch korrekt durchgeführt worden?
- Können die Schlußfolgerungen der Evaluation durch die gewonnen Daten ausreichend begründet werden?
- Sind während der Evaluation Sicherheitsmaßnahmen getroffen worden, damit die Ergebnisse nicht durch persönliche Gefühle und Vorurteile der Evaluatoren verfälscht werden?

3. Zusammenfassung.

Die oben in Gruppen zusammengefaßten „Sünden“ von Evaluationen zeigen recht deutliche Unterschiede zu den Bewertungskriterien für Grundlagen- und angewandten Wissenschaften:
- *Grundlagenwissenschaft*: Hier hat die Merkmalsgruppe der Objektivität die ausschlaggebene Priorität. Sowohl Anwendbarkeit und Durchführbarkeit sind untergeordnet, moralische öberlegungen gibt es keine.
- *Angewandten Wissenschaften:* Neben intersubjektive Überprüfbarkeit,

haben auch Anwendbarkeit und Durchführbarkeit ihre Bedeutung. Ethische Kritierum jedoch kaum.
- *Evaluationen*: Hier haben alle vier Merkmalsgruppen gleichrangige Bedeutung.

Ess zeigt sich hier als herausragende Besonderheite von Evaluationen der Umgang mit Werten und damit auch der Umgang mit ethisches Problemen. Ethische Fragestellungen sind jedoch nicht von entsprechenden Vorstellungen von Gerechtigkeit (bzw. Theorien der Gerechtigkeit) zu trennen. Nach dem bisher Gesagten ist nun wohl deutlich geworden, welch enorme Bedeutung ethische Überlegungen für die Evaluationstheorie und -praxis haben.

Ohne hier näher auf diese Fragen eingehen zu können (vgl. dazu zum Thema Evaluation vor allem House, 1980 und ganz allgemein vor allem Rawls 1990), möchte ich abschließen noch verschiedene Dimensionen der Gerechtigkeit auflisten:

- *Angebot* für alle gleich
- *Zugang* für alle gleich
- *Teilnahme/Inanspruchnahme* für alle Gruppen gleich
- *Erreichbarkeit* (attainment)/*Erfolgsquote* für alle Gruppen gleich
- *Fertigkeiten/Leistungen* (proficiency) für alle Gruppen gleich
- *Ersehnbarkeit* (aspiration)/*Bedürfnisse* für alle Gruppen gleich
- *Auswirkungen* (impacts) für alle Gruppen gleich

Voilá! Hier sind sie nochmals, die 10 Todsünden der Medienevaluation:

(1) Reduktion auf Daten
(2) Reduktion auf Methodenlehre
(3) Reduktion auf angewandte Sozialforschung
(4) Zielorientierte Ansätze ohne Hinterfragung der Wertehierarchie
(5) Quasi-empirisches Forschungsdesign
(6) Akzeptanzstudien
(7) Kategorienfehler
(8) Fehler bei der Skalenzuordnung
(9) Gewichtungsfehler
(10) Gerechtigkeits/Fairnessfehler

Literatur

Bateson, G. 1972. Steps to an Ecology of Mind. A Revolutionary Approach to Man's Understanding of Himself. New York: Ballantine Books.

Baumgartner, P. 1993. *Der Hintergrund des Wissens. Vorarbeiten zu einer Kritik der programmierbaren Vernunft*. Klagenfurt: Kärntner Druck- und Verlagsanstalt.

Baumgartner, P. 1995. Didaktische Anforderungen an (multimediale) Lernsoftware. In: *Information und Lernen mit Multimedia*, Hg. von L. J. Issing und P. Klimsa. Weinheim: Psychologie- Verl.-Union. 241-252.

Baumgartner, P. und S. Payr. 1996. Learning as action: A social science approach to the evaluation of interactive media. In: *Proceedings of ED-MEDIA 96 - World Conference on Educational Multimedia and Hypermedia*, Hg. von P. Carslon und F. Makedon. Charlottesville: AACE. 31- 37.

Baumgartner, P. 1997. Evaluation vernetzten Lernens: 4 Thesen. In: Virtueller Campus. Forschung und Entwicklung für neues Lehren und Lernen, Hg. von H. Simon. Münster: Waxmann. 131- 146.

Baumgartner, P. und S. Payr. 1997. Methods and practice of software evaluation: The case of the European Academic Software Award (EASA). In: Proceedings of ED-MEDIA 97 - World Conference on Educational Multimedia and

Hypermedia. Charlottesville: AACE. 44-50.

Baumgartner, P. 1999. Evaluation mediengestützten Lernens. Theorie - Logik - Modelle. Erscheint in: ???, Hg. von ??? Münster: Waxmann.

Biermann, H. 1994. Lehren und Lernen mit Computern. In: *Lehren und Lernen im Umfeld neuer Technologie: Reflexionen vor Ort*, Hg. von J. Petersen und G.-B. Reiner. Frankfurt/Main: Peter Lang. 123-141.

Doll, C. A. 1987. *Evaluating Educational Software*. Chicago/London: American Library Association.

Fricke, R. 1995. Evaluation von Multimedia. In: *Information und Lernen mit Multimedia*, Hg. von L. J. Issing und P. Klimsa. Weinheim: Psychologie-Verl.-Union. 401-413.

Habermas, J. 1981. Theorie des kommunikativen Handelns. Handlungsrationalität und gesellschaftliche Rationalisierung. Frankfurt/M.: Suhrkamp.

Habermas, J. 1981. Theorie des kommunikativen Handelns. Zur Kritik der funktionalistischen Vernunft. Frankfurt/M.: Suhrkamp.

Habermas, J. 1984. Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns. Frankfurt/M.: Suhrkamp.

House, E. R. 1980. *Evaluating with validity*. Beverly Hills, CA: SAGE.

Neuweg, G. H. 1998. Implizites Wissen. Eine Studie zur berufspädagogischen Bedeutung der Erkenntnis- und Wissenstheorie Michael Polanyis, Fakultät für Sozial- und Wirtschaftswissenschaften, Abteilung für Berufs- und wirtschaftspädagogik, Universität Linz.

Polanyi, M. 1962. Personal Knowledge. Towards a Post-Critical Philosophy. Chicago/London: Chicago Press.

Polanyi, M. 1969. Knowing and Being. Essays edited by Marjorie Grene. Chicago/London: Chicago Press.

Rawls, J. 1990. Eine Theorie der Gerechtigkeit. 5. Aufl. Frankfurt/Main: Suhrkamp.

Ryle, G. 1969. Der Begriff des Geistes. Stuttgart: Reclam.

Scriven, M. 1991a. *Evaluation Thesaurus*. 4. Aufl. Newbury Park: SAGE.

Scriven, M. 1991b. Introduction: The Nature of Evaluation. In: *Evaluation Thesaurus*, Hg. von M. Scriven. 4. Aufl. Newbury Park: SAGE. 1-43.

Shadish, W. R. J., T. D. Cook und L. C. Leviton. 1991. *Foundation of Program Evaluation. Theories of Practice*. Newbury Park: SAGE.

Stufflebeam, D. L. und A. J. Shinkfield. 1985. *Systematic Evaluation*. Boston: Kluwer.

Thomé, D. 1988. *Kriterien zur Bewertung von Lernsoftware*. Heidelberg: Hüthig.

Thorndike, R. M., G. K. Cunningham, R. L. Thorndike et al. 1991. *Measurement and Evaluation in Psychology and Education*. 5. Aufl. New York: Macmillan.

Weber, M. 1988a. Der Sinn der "Wertfreiheit" der soziologischen und ökonomischen Wissenschaften. In: *Gesammelte Aufsätze zur Wissenschaftslehre*, Hg. von M. Weber. 7. Aufl. Tübingen: UTB Mohr. 489-540.

Weber, M. 1988b. Die "Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In: *Gesammelte Aufsätze zur Wissenschaftslehre*, Hg. von M. Weber. 7. Aufl. Tübingen: UTB Mohr. 146-214.

Wiersma, W. 1991. *Research Methods in Education*. 5. Aufl. Needham Heights, MA: Simon & Schuster.

Worthen, B. R. und J. R. Sanders. 1987. *Educational evaluation: Alternativ approaches and practical guidelines*. White Plains: Longman.