Tergan, Sigmar-Olaf

Qualitätsbeurteilung von Bildungssoftware mittels Kriterienkatalogen. Problemaufriss und Perspektiven

Unterrichtswissenschaft 29 (2001) 4, S. 319-341

Unterrichtswissenschaft

Zeitschrift für Lernforschung
29. Jahrgang / 2001 / Heft 4

Thema:
Lehren und Lernen mit multimedialen Lernumgebungen

Verantwortlicher Herausgeber:
Günter Dörr

Sigmar-Olaf Tergan

Qualitätsbeurteilung von Bildungssoftware mittels Kriterienkatalogen. Problemaufriss und Perspektiven

Checklists for Evaluation of Educational Software – Problems and Perspectives

Kriterienkataloge erfreuen sich bei der Qualitätsbeurteilung von Bildungssoftware nach wie vor einer großen Beliebtheit. Dies gilt sowohl für die Evaluation von Offline- als auch für Online-Bildungsangebote. Ihre Anwendung wird jedoch aufgrund bestehender Probleme und Schwächen zunehmend kritisiert. Es können drei Problemtypen unterschieden werden: Konzeptuelle Mängel (z.B. Übergewicht technischer und eindimensionaler Kriterien), Schwächen bezüglich Gütekriterien (z.B. mangelnde Beurteilerübereinstimmung, unklare prädiktive Validität) und Anwendungsprobleme (z. B. mangelnde Gerichtetheit und Anpassbarkeit von Kriterien an unterschiedliche Programmtypen und situative Rahmenbedingungen). Im vorliegenden Beitrag werden bestehende Probleme aufgezeigt und diskutiert. Abschließend werden Perspektiven für den effektiven Einsatz von Kriterienkatalogen bei der Qualitätsbeurteilung von Bildungssoftware dargestellt.

Checklists for the evaluation of educational software are still very popular. This is true for the evaluation of offline- as well as online-software. However, checklists are increasingly criticized because of inherent shortcomings and existing problems in application. One may distinguish three problem types: conceptual problems (e.g. preponderance of technical and uni-dimensional criteria), problems with respect to test-statistical criteria (e.g. lack of reliability and unknown validity of ratings), and application problems (e.g. lack of tailored criteria and adaptivity of criteria with respect to different program types and contextual constraints). The paper outlines and discusses in detail existing problems and shortcomings. Perspectives for an effective use of checklists under conditions of different contextual constraints are outlined.

1. Einleitung

Angesichts der zunehmenden Bedeutung computergestützten Off- und Online-Lernens, der Vielfalt existierender Software-Anwendungen und den Problemen von Käufern und Anwendern beim Vergleich und der Auswahl von Bildungssoftware besteht ein großer Bedarf an leicht handhabbaren, ökonomischen und vielseitig verwendbaren Verfahren zur Qualitätsbeurteilung. Diesem Bedarf versuchen sogenannte Kriterienkataloge gerecht zu werden. Kriterienkataloge enthalten Zusammenstellungen von Fragen und Einschätzungsskalen zur standardisierten Beschreibung und Beurteilung von Aspekten der technischen, inhaltlichen und didaktischen Qualität von Bildungssoft-

ware. Grundlage der Qualitätsbeurteilung sind sogenannte Qualitätskriterien. Diese thematisieren die von den Katalogentwicklern für die Produktqualität als bedeutsam erachteten Merkmale, z.B.: Textgestaltung, Anwenderfreundlichkeit, Bildschirmaufbau, Interaktivität. Fricke (2000) beschreibt Qualitätskriterien aus wissenschaftlicher Perspektive. Er bezeichnet sie als „... allgemeine Merkmale einer Lernsoftware, deren Lernwirksamkeit in einer Validitätsstudie wissenschaftlich nachgewiesen wurden. Den Begriff 'Qualitätskriterium' verwendet man in der Praxis allerdings schon dann, wenn lediglich Vermutungen über die Lernwirksamkeit eines Programmmerkmals vorliegen" (S. 75). Von Qualitätskriterien spricht man ferner dann, wenn es um die Einhaltung bestimmter Standards geht. So beziehen sich beispielsweise Qualitätskriterien zur Beurteilung von Aspekten der Softwareergonomie an entsprechenden Vorgaben internationaler Normierungsbemühungen wie der ISO 9241, einem umfassenden Regelwerk für ergonomische Anforderungen an rechnergestützte Büroarbeit (vgl. Willumeit, Gediga & Hamborg (1996). Kriterienkataloge verlangen in der Regel eine systematische Einschätzung einer Software nach den vorgegebenen Kriterien. Über eine Zusammenfassung von zum Teil gewichteten Einzelurteilen gelangen Anwender zu einem Gesamturteil bezüglich der Softwarequalität (z.B. Gräber, 1990; Meier, 1995).

Eine Übersicht über bekannte deutsche und englischsprachige Kriterienkataloge geben Meier (1995) und Gräber (1996). Theoretische und empirische Arbeiten zur Anwendung von Kriterienkatalogen für die Qualitätsevaluation von Lern- und Informationssystemen enthält das Buch von Schenkel, Tergan & Lottmann (2000). Hier werden u.a. Ergebnisse einer detaillierten Analyse der Stärken und Schwächen ausgewählter deutschsprachiger Kriterienkataloge bei der Qualitätsbeurteilung multimedialer Lern- und Informationssysteme mitgeteilt. Eine Analyse der Leistungsfähigkeit und Übereinstimmung entsprechender Kriterienkataloge bei der Qualitätsbeurteilung multimedialer Lern- und Informationssysteme wurde von Tergan (2000 c) vorgenommen.

Kriterienkataloge eignen sich für eine differenzierte Beschreibung von Softwaremerkmalen und können damit Anwendern bei der Entscheidung für den praktischen Einsatz einer Bildungssoftware im Unterrichtskontext behilflich sein. Diese Funktion hat beispielsweise die SODIS-Datenbank (http://www.sodis.de/). Die Datenbank wurde vom Landesinstitut für Schule und Weiterbildung in Paderborn zur Erleichterung der Selektion und Auswahl von Bildungssoftware angelegt. Grundlage der Evaluation ist das Software-Dokumentations- und -Informationssystem (SODIS). SODIS ist ein kriterienkatalog-orientiertes Verfahren für die Qualitätsbeurteilung von schulischer Bildungssoftware. Eine Beschreibung des Verfahrens und seiner praktischen Anwendung gibt Korbmacher (2000). Kriterienkataloge eignen sich durch die Einbeziehung sogenannter K.o.-Kriterien, z.B. Kriterien wie „Absturzsicherheit eines Lernprogramms", insbesondere für eine schnelle und kostengünstige Vorselektion von Bildungssoftware (Fricke, 2000). Die Erwartung von Entwicklern und Anwendern von Kriterienkatalogen besteht

darüber hinaus häufig darin, dass bei Durchführung der Qualitätsbeurteilung einer gegebenen Bildungssoftware mittels eines Kriterienkataloges auf weitere Evaluationsmaßnahmen, insbesondere auf eine empirische Überprüfung der Wirkungen dieser Software verzichtet werden kann. Diese Erwartung gründet auf der Vorstellung eines vollständigen, objektiven und prognostisch validen Instrumentariums. Eine weitere verbreitete Auffassung besteht darin, dass Kriterienkataloge leicht handhabbar sind und mit ihrer Hilfe eine Qualitätsbeurteilung auf dem Markt verfügbarer Bildungssoftware auch von Nicht-Experten in einer Art Warentest durchgeführt werden kann.

Die Anwendung von Kriterienkatalogen zur Qualitätsbewertung zwecks Auswahl geeigneter Software für bestimmte Anwendungsziele ist nur eine unter mehreren Verwendungsmöglichkeiten. Andere Verwendungsmöglichkeiten im Kontext eines umfassenden Qualitätssicherungsprozesses bestehen bei der Entwicklung von Bildungssoftware (Tergan, 2000 a; Schenkel, 2000). Kriterienkataloge können im Rahmen der Qualitätssicherung eine wichtige Funktion erfüllen, indem sie Bewertungsaspekte für unterschiedliche Auflösungsgrade einer Qualitätsevaluation präzisieren und entsprechende Bewertungskriterien bereitstellen. Je nachdem, in welcher Phase des Qualitätssicherungsprozesses Kriterienkataloge eingesetzt werden, kann ihre Funktion unterschiedlich sein (vgl. Tergan, 2000 a). Im Rahmen formativer Evaluation besteht ihre Funktion darin, Kriterien im Sinne von Leitlinien für die Softwaregestaltung im Rahmen von Planungs-, Entwicklungs- und Prototyping-Aktivitäten bereitzustellen und Entwicklungsprozesse transparenter zu machen. In der Phase der summativen Evaluation können Kriterienkataloge in Ergänzung zu anderen Methoden, zum Beispiel Adressatenbefragungen und empirischen Methoden, bei der Beurteilung der pädagogischen Effektivität eingesetzt werden. Die ergänzende Anwendung von Kriterienkatalogen in dieser Phase kann dazu dienen, jene Design-Merkmale der evaluierten Varianten sowie Gemeinsamkeiten und Unterschiede offenzulegen, die zum Beispiel für eine systematische theorieorientierte Variation von Bedingungen im Rahmen eines experimentellen Vorgehens und für die Theorieentwicklung hilfreich sein können (vgl. Fezzardi, Hasebrook & Glowalla, 1992).

2. Probleme und Schwächen von Kriterienkatalogen

Kriterienkataloge gehören zum Standardrepertoire an Methoden, die für die Qualitätsbeurteilung von Bildungssoftware verwendet werden. Ihre Anwendung ist jedoch nicht unproblematisch. Auf einige zentrale Probleme und Schwächen wird in der Literatur vereinzelt hingewiesen (u.a. Baumgartner, 1995; Fricke, 2000; Prichard, Micceri & Barrett 1989; Squires & McDougall, 1994; Tergan, 1998). Dies sind zum einen allgemeine Probleme, die mit der Konstruktion und Anwendung von Fragebogen sowie mit der Auswertung und Interpretation von Skalenwerten als Ergebnisse subjektiver Einschätzungen verbunden sind (u.a. Jacobs, 1998). Auf diese wird im Rahmen des vorliegenden Beitrags nicht näher eingegangen. Es sind zum anderen konzeptuelle, inhaltliche, formale und praktische Probleme, die die Funktio-

nalität und Effektivität von Kriterienkatalogen für die Qualitätsbeurteilung von Bildungssoftware betreffen. Letzteren Problemen und Möglichkeiten zu deren Lösung gelten die Erörterungen dieses Beitrags.

Im Folgenden werden die bestehenden Probleme und Schwächen von Kriterienkatalogen umfassend und systematisiert dargestellt und diskutiert. Es werden drei Problemtypen unterschieden: Konstruktionsmängel (z.B. Übergewichtung technischer Kriterien, Dominanz eindimensionaler Kriterien), Schwächen bezüglich teststatistischer Gütekriterien (z.B. mangelnde Beurteilerübereinstimmung, unklare prädiktive Validität) und Anwendungsprobleme (z. B. mangelnde Gerichtetheit und Anpassbarkeit von Kriterien an unterschiedliche Programmtypen und situative Rahmenbedingungen).

2.1 Konzeptuelle Schwächen

Unschärfe des Begriffs „Qualitätskriterium"

Für die überwiegende Mehrzahl der Kriterienkataloge ist eine unscharfe Verwendung des Begriffs „Qualitätskriterium" kennzeichnend. Der Begriff „Qualitätskriterium" im Zusammenhang mit einem Softwaremerkmal sollte nach Fricke (2000) eigentlich auf ein Merkmal verweisen, dessen Lernwirksamkeit in einer Validitätsstudie wissenschaftlich nachgewiesen wurde. Nicht zuletzt mangels ausreichender empirischer Belege verwendet man den Begriff in der Praxis der Softwarebeurteilung allerdings schon dann, wenn von einem Kriterium „aufgrund Erfahrung, plausibler Schlüsse etc. vermutet werden kann, dass das betreffende Merkmal das Lernen positiv beeinflusst" (Fricke, 2000, S. 75).

Probleme der Standardisierung

Kriterien bedürfen der Operationalisierung, um ihre standardisierte Anwendung im Rahmen von Kriterienkatalogen zu gewährleisten. Die in Kriterienkatalogen verwendeten Kriterien sind jedoch selten standardisiert. Es fehlen in der Regel Angaben darüber, wann bei einer Bildungssoftware ein bestimmtes Kriterium als erfüllt oder in Teilen erfüllt gelten soll. Diese konzeptuelle Schwäche hat Auswirkungen auf die Güte der Beurteilung. Wenn nicht eindeutig klar ist, welche einzelnen Aspekte in die Beurteilung der Erreichung eines Kriteriums einzubeziehen sind, geht dies zu Lasten der Objektivität und führt in der Praxis zu einer Reduzierung der Beurteilerübereinstimmung.

Übergewicht technischer Kriterien

In der Mehrzahl der vorliegenden Kriterienkataloge werden technische Merkmale der Software und Kriterien übergewichtet (Squires & McDougall, 1994; Tergan, 1998; Fricke, 2000). Fricke stellt hierzu fest: „Leider überwiegen in den bekannten Katalogen (vgl. Meier, 1995) ... Kriterien hauptsächlich aus dem Bereich der 'Gestaltung der Benutzeroberflächen' und 'Technik des Programmablaufs'" (S. 75). Bestimmte technische Merk-

male und deren Vorhandensein stellen zwar Voraussetzungen für eine technisch einwandfreie und effektive Nutzbarkeit von Bildungssoftware dar. Technische Mängel (z.B. Programmabstürze) bzw. das Fehlen entsprechender technischer Voraussetzungen (z.B. fehlende Lauffähigkeit auf gängigen Plattformen) können auch K.o.-Kriterien bei der Beurteilung der Verwendbarkeit von Bildungssoftware darstellen. Technische Merkmale von Medien sind jedoch insgesamt gesehen keine guten Prädiktoren von Lernleistungen (Clark & Craik, 1992; Kerres, 1998). Diese Einschätzung gilt auch für Lernprogramme (Hasebrook, 1995; Schulmeister, 1996). Weitaus wichtigere Prädiktoren sind häufig Lernvoraussetzungen auf Seiten der Nutzer sowie instruktionale, curriculare und situative Rahmenbedingungen des Software-Einsatzes (Kerres, 1998; vgl. auch Zimmer & Psaralidis, 2000).

Dominanz eindimensionaler Items

Es besteht eine Dominanz von Fragen zu Einzelmerkmalen einer Bildungssoftware. Fragen, die das Vorkommen bestimmter hardware- oder softwaretechnischer Merkmale betreffen, sind in der Regel eindimensional. Gefragt wird: Ist ein bestimmtes Systemmerkmal vorhanden? Wie verständlich ist der verwendete Text? Liegen Aufgaben zum Selbsttesten vor? Auch hinsichtlich pädagogisch-didaktischer Softwaremerkmale wird in Kriterienkatalogen vor allem eindimensional gefragt. Zum Beispiel: Wird eine bestimmte Lehrfunktion unterstützt? Mehrdimensionale Fragen sind in Kriterienkatalogen eher selten. Mehrdimensionale Fragen thematisieren Zusammenhänge. Gefragt wird zum Beispiel: Entspricht das Niveau der Darstellung des Sachverhalts den Lernvoraussetzungen der Adressaten? Sind die eingesetzten didaktischen Mittel für eine Unterstützung von Adressaten mit unterschiedlichen Lernvoraussetzungen funktional?

Das Vorhandensein wesentlicher Hard- und Softwaremerkmale kann zwar mittels eindimensionaler Fragen ökonomisch abgefragt und Software im vergleichenden Bewertungsverfahren nach K.o.-Kriterien ausgesondert werden. Der Sachverhalt, dass in Kriterienkatalogen bevorzugt eindimensionale Fragen Verwendung finden, diese isoliert betrachtet werden (Baumgartner, 1995) und bestehende Wechselwirkungen zwischen lernrelevanten Variablen (Vorwissen, Inhaltskomplexität, situative Bedingungen) in entsprechenden Kriterien kaum zum Tragen kommen, hat jedoch einen negativen Einfluss auf die prädiktive Validität von Kriterienkatalogen. Denn, die Wirksamkeit von Merkmalen einer Bildungssoftware äußert sich nicht selten erst in Wechselwirkung mit Lernervariablen und situativen Bedingungen des Softwareeinsatzes (z.B. Fricke, 1991; 2000). Andererseits besteht das Dilemma, dass es bei systematischer Berücksichtigung möglicher Wechselwirkungen zwischen Softwaremerkmalen, Lernvoraussetzungen und situativen Bedingungen des Softwareeinsatzes zu einer kombinatorischen Explosion der einzubeziehenden Kriterien kommen könnte. Eine Orientierung der Kriteriengenerierung an theoretischen Modellen und eine die Wechselwirkungen zwischen lernrelevanten Variablen berücksichtigende, eher verstehensorientierte ganzheitliche Evaluation ("comprehensive evaluation") erweist sich

als unerlässlich und könnte einer derartigen Entwicklung Grenzen setzen (vgl. Squires & McDougall, 1996; Tergan, 1998; Schott, 2000).

Mangelnde Berücksichtigung von Kriterien für selbstgesteuertes Lernen

Die überwiegende Mehrzahl der bestehenden Kriterienkataloge (vgl. Meier, 1995) fokussiert vorrangig traditionelle Lernprogramme wie Übungsprogramme und tutorielle Systeme entsprechend dem traditionellem kognitiven Paradigma angeleiteten Lernens, seltener Simulationen und Simulationsumgebungen sowie multimediale und hypermediale Bildungssoftware für das selbstgesteuerte Lernen (vgl. jedoch Tolhurst, 1992). Für die Qualitätsbeurteilung von Simulations-Lernumgebungen, multi- und hypermedialer Software, hybrider sowie telematischer Lernumgebungen sind entsprechende Kataloge nur bedingt geeignet (vgl. Bangert-Drowns & Kozma, 1989; Jacobson & Spiro, 1994; Tergan, 1998). So erfordert die Evaluation von hypermedialer Offline- und Online-Software eine besondere Berücksichtigung von Gestaltungs- und Nutzungskriterien bezüglich der Benutzeroberfläche sowie weitere spezifische Kriterien, die den Besonderheiten selbstgesteuerten, konstruktivistischen Lernens gerecht werden (Tergan, 1998). Bei Online-Angeboten wären zusätzlich Merkmale der Website sowie unterschiedlicher Formen betreuten Tele-Lernens sowie bestehende synchrone und/oder asynchrone Kommunikations- und Kooperationsmöglichkeiten zu berücksichtigen (z.B. Astleitner, 1997; Kerres, 1998; Friedrich & Hron, im Druck).

Korbmacher (2000) kritisiert in diesem Sinne das Software-Dokumentations- und -Informationssystem SODIS. Bei der Anwendung von SODIS im Rahmen der Evaluation des für die Weiterbildung Erwachsener entwickelten multimedialen Informations- und Lernprogramms Informations- und Kommunikationstechniken im Handwerk (IKTH) wurden Schwächen deutlich, die für schulorientierte Kriterienkataloge typisch sind: Korbmacher beschreibt diese folgendermaßen: „Das Kriterienraster ist auf schulische Lernprozesse zugeschnitten. ... Eine Software, die ausschließlich für das Selbststudium gedacht ist, kann deshalb m.E. mit Hilfe des gegenwärtigen SODIS-Rasters nicht adäquat beurteilt werden" (S. 216).

Meier (2000) äußert sich in einer zusammenfassenden Beurteilung der Eignung der kriterienorientierten Beurteilungsverfahren MEDA und MEDA´97 in ähnlicher Weise. Sie stellt im Hinblick auf die Qualitätsbeurteilung des Lern- und Informationssystems „Informations- und Kommunikationstechniken im Handwerk – IKTH" (Lottmann, 2000) fest, dass die Verfahren bei einer strikten Anwendung auf der Ebene der sog. Anwendungsanalyse zu einer unangemessenen Qualitätsbeurteilung führen. Nach Meier sind bei den Verfahren „die meisten der Fragen und Kriterien, die den hier formulierten Analyseaspekten zugewiesen sind, auf reine Lernprogramme zugeschnitten, die in herkömmlichen Unterricht zu integrieren sind und bei deren Bearbeitung ein Ausbilder zugegen ist" (S. 179). Bei Softwareprodukten wie IKTH, die ausdrücklich als Selbstlern- und Informationsprogramm konzipiert sind, „müssen die meisten der von MEDA und MEDA´97 formulierten Fragen verneint werden" (S. 179). Bei einem starren Schema der Datenaggregation zur

Ermittlung eines allgemeinen Qualitätsurteils können sich hierdurch Urteilsverzerrungen ergeben. Da MEDA'97 die Möglichkeit vorsieht, aus einem Angebot an Fragen zur Qualitätsbeurteilung irrelevante Fragen auszuschließen bzw. fehlende hinzuzufügen, hält Meier das Beurteilungsinstrument MEDA dennoch für „ein geeignetes Werkzeug, bestimmte Schwachpunkte auch aus ‚Mischprogrammen' ... zu eliminieren und zu analysieren"... „Dem Beurteiler bleibt die Aufgabe, in der Analyse festgestellte ‚Ausreißer' gemessen an Intention und Konzeption der zu analysierenden Software eine Wertigkeit in der gesamten Beurteilung der Qualität zuzuweisen" (Meier, 2000, S. 184). In der Bewertung von Meier wird deutlich, dass der Expertise des Beurteilers eine große Bedeutung beigemessen wird.

Mangelnde Kontextbezogenheit

Der Zweck der Entwicklung und Anwendung von Kriterienkatalogen wird vielfach darin gesehen, die Qualität von Bildungssoftware als solche, d.h. unabhängig von den Lerninhalten und dem Kontext ihrer Anwendung beurteilen zu können (Baumgartner & Payr (1997). Kontextbezogene Kriterien wie z.B die curriculare Integration einer Bildungssoftware oder die tutorielle Betreuung der Lernenden bei der Softwareanwendung als besondere situative Bedingungen des Softwareeinsatzes bleiben unberücksichtigt. Eine kontextsensitive Evaluation von Bildungssoftware, wie sie aufgrund der Erkenntnisse zur Kontextgebundenheit menschlichen Lernens erforderlich wäre, kann damit nicht geleistet werden (vgl. Squires & McDougall (1996). Für konstruktivistisch orientierte sowie problemorientierte Lernumgebungen, in denen die Kontextorientierung eine große Rolle spielt, fehlen in aller Regel entsprechende Beurteilungskriterien.

Theoretische Orientierungslosigkeit

Fricke (2000) kritisiert an den gegenwärtig bestehenden Kriterienkatalogen, dass diese selten aus Lehr-Lerntheorien abgeleitete Kriterien enthalten und sich nicht an bestehenden Instruktionsdesign-Modellen orientieren. Baumgartner (1995) verweist auf die theoretische Orientierungslosigkeit von Kriterienkatalogen. "Die Erstellung umfangreicher und detaillierter Kriterienkataloge ohne allgemein akzeptierte Gewichtungsverfahren vernachlässigt die Frage der zugrunde liegenden Lerntheorie. Damit wird aber der eigentliche Sinn von Kriterienkatalogen unterlaufen: Vor lauter Bäumen (Kriterien) wird der Wald (pädagogische und didaktische Angemessenheit) nicht mehr gesehen. Zusätzlich besteht die Gefahr, dass durch die isolierte Betrachtung der Lernprogramme ihre didaktische Einbindung in ein Curriculum und die ganzheitliche Gestaltung der Lernsituation vernachlässigt wird" (S. 242). Es besteht die Notwendigkeit, Kriterienkataloge in enger Orientierung an theoretischen Konzeptionen zu entwickeln, diese offenzulegen und Beurteiler in die Lage zu versetzen, Beurteilungen von Bildungssoftware entsprechend der jeweils präferierten Konzeption vorzunehmen. Ansätze zur Überwindung der theoretischen Orientierungslosigkeit werden bei Fricke (1995; 2000), Schott (2000) und Tergan (1998) diskutiert.

Mangelnde Ganzheitlichkeit der Qualitätsevaluation

Schott (2000) kritisiert eine mangelnde Ganzheitlichkeit der Qualitätsevaluation mittels Kriterienkatalogen. Er sieht die Gefahr, „dass die Urteile zu den einzelnen Kriterien additiv zusammengetragen werden und eine ganzheitliche Sicht behindern" (S. 108). Eine ganzheitliche Evaluation sei notwendig, „weil die Güte einer Bildungsmaßnahme ... vom Gesamtzusammenhang der sie beeinflussenden Faktoren abhängt" (S. 108). Solche Faktoren sind nach Schott unterschiedliche Eigenschaften der Lernenden, des Lehrstoffs, der Lehrmethoden und Medien sowie Rahmenbedingungen des Umfeldes, in dem die Bildungsmaßnahme stattfindet (vgl. Tergan, Hron & Mandl, 1992; Tergan, 1998). Bei der Beurteilung der pädagogischen Qualität von Bildungssoftware für den Einsatz in einem bestimmten Unterrichtskontext oder für die private Nutzung bleiben Rahmenbedingungen jedoch nicht zuletzt infolge fehlender Informationen üblicherweise unberücksichtigt. Hierunter leidet die Ganzheitlichkeit der Qualitätsevaluation.

2.2 Schwächen bezüglich Gütekriterien

Geringe Beurteilerübereinstimmung

Für einen erfolgreichen Einsatz von Kriterienkatalogen, z.B. bei der Programmauswahl, ist eine hinreichende Übereinstimmung von Beurteilern bei der Einschätzung der Qualität einzelner Softwareaspekte (Kriterien) sowie der Gesamtqualität der Software eine wichtige Voraussetzung. Ist die Beurteilerübereinstimmung gering, so mangelt es dem betreffenden Verfahren an Objektivität. Geringe Objektivität der Urteile bei den einzelnen Kriterien hat negative Auswirkungen auf die prädiktive Validität. Ingenkamp (1995, S. 37) stellt hierzu fest: „Wer auf Objektivität verzichtet, gibt auch Zuverlässigkeit und Validität auf". Kriterienkataloge bieten zwar durch die Vorgabe von Kriterien, Einschätzungsskalen etc. Möglichkeiten zu einer standardisierten Erhebung von Qualitätsurteilen. Eine hinreichende Objektivität der Beurteilungen könnte von daher erwartet werden. Ergebnisse empirischer Untersuchungen zeigen jedoch, dass bei Kriterienkatalogen nicht selten eine geringe Beurteilerübereinstimmung festzustellen ist. So berichten Jolicoer & Berger (1986), die in einer Metaanalyse die Übereinstimmung der Qualitätsbewertungen bei 82 Programmen zweier amerikanischer Institutionen (Educational Products Information Exchange, Microsift) untersuchten, nur eine geringe Beurteilerübereinstimmung. Prichard, Micceri & Barrett (1989) stellen fest, dass unterschiedliche Beurteiler bei der Anwendung eines Kriterienkataloges zu recht unterschiedlichen Urteilen gelangten und ein hohes Maß an Beurteilertraining notwendig war, um eine hinlängliche Urteilsübereinstimmung zu erzielen. Entsprechende Ergebnisse neuerer Untersuchungen deuten in dieselbe Richtung (Fricke, 2000).

Bei der Qualitätsbeurteilung durch unterschiedliche Experten kommen offensichtlich unterschiedliche individuelle Beurteilungstendenzen zum Tragen, die die Ursache für die geringe Objektivität der Qualitätsurteile darstellen. So

berichtet Tergan (2000 b) von den Ergebnissen einer Untersuchung, in der Experten befragt wurden, welche Bedeutung sie bestimmten Qualitätsaspekten für die Beurteilung der Qualität von Bildungssoftware beimessen. „Es konnte gezeigt werden, dass sich Experten hinsichtlich der Einschätzung von Qualität, Funktionalität und Nutzen von Bildungssoftware sowie einzelner Softwaremerkmale zum Teil erheblich voneinander unterscheiden. Die unterschiedlichen Bewertungen einzelner Softwaremerkmale durch verschiedene Experten legen nahe, dass Unterschiede bezüglich individueller Annahmen über die Bedeutung des Einflusses einzelner Softwaremerkmale auf die pädagogische Effektivität einer Lern- und Informationssoftware bei Verwendung von Experten-Beurteilungs-Verfahren, wie zum Beispiel Kriterienkatalogen, zu mangelnder Beurteilerübereinstimmung beitragen" (S. 159).

Möglichkeiten zur Verbesserung der Urteilsübereinstimmung und damit von Objektivität und Zuverlässigkeit von Kriterienkatalogen bieten sich, indem die Kriterien hinsichtlich ihrer Bedeutung in differenzierter Weise – z.B. unter Rückgriff auf theoretische Annahmen und didaktische Modelle unter Verwendung von Positiv- und Negativ-Beispielen der Software-Gestaltung – erläutert werden. Sofern einschlägige empirische Befunde vorliegen, sollten diese in die Erläuterungen einbezogen werden. Um die Objektivität und Vergleichbarkeit der Messungen zu erhöhen, empfiehlt es sich, wann immer dies sinnvoll und möglich ist, Software-Merkmale quantitativ zu beschreiben. Beispiele hierfür werden bei Prichard et al. (1989), Gräber (1990) und Meier (1995) beschrieben. Ratingskalen dienen vorrangig der erleichterten Auswertung von Merkmalseinschätzungen. Sie erweisen sich als geeignet, um Zusammenfassungen von Einzelurteilen zu einem Merkmalsbereich sowie Vergleiche zwischen den Urteilen verschiedener Beurteiler zu erleichtern und damit ein Beurteilertraining besser durchführen und Ergebnisse eines entsprechenden Trainings leichter kontrollieren zu können (vgl. AKAB in Meier, 2000).

Kriterienbeschreibungen helfen zwar, den Geltungsbereich eines Kriteriums zu umschreiben. Sofern sie allerdings nicht auch begründen, unter welchen Bedingungen einer antizipierten Anwendung von Bildungssoftware welchem Kriterium welcher Ausprägungsgrad bei welchen Zielen des Lernens beizumessen ist, bleiben alle Entscheidungen dem Beurteiler überlassen. Diese haben jedoch nicht selten unterschiedliche, auf individuellen Erfahrungen und Bewertungsmustern gründende Vorstellungen bezüglich Qualitätsmerkmalen von Bildungssoftware, die einer Verbesserung der Objektivität eines Verfahrens entgegenstehen und dessen prädiktive Validität beeinflussen (Tergan, 2000 b). Ein gezieltes Beurteilertraining sowie die ergänzende Anwendung empirischer Verfahren wie z.B. Tests oder Fallstudien ist daher empfehlenswert (vgl. Joliceur & Berger, 1986; Jacobs, 1998; Tergan, 1998; Schenkel et al., 2000).

Unklare Validität und Gewichtung der Kriterien

Eine Voraussetzung des erfolgreichen Einsatzes von Kriterienkatalogen bei der Evaluation von Bildungssoftware ist die Konstrukt-Validität und die prä-

diktive Validität der einbezogenen Kriterien. Zu fragen ist: Erfasst ein bestimmtes Kriterium tatsächlich einen bedeutsamen Aspekt der pädagogischen Qualität? Steht das so erfasste Merkmal im Zusammenhang mit dem Lernerfolg der Softwarenutzer? Ein Merkmal kann dann als prognostisch valide angesehen werden, „wenn zumindest eine Korrelation zwischen dem Ausprägungsgrad des Merkmals und dem Lernergebnis vorhanden ist" (Frike, 2000, S. 75).

Die Validität der Beurteilung der technischen Qualität einer Bildungssoftware mit Hilfe von Kriterienkatalogen ist aufgrund der allgemein bekannten technischen Standards im Hard- und Softwarebereich in aller Regel gewährleistet. Die prädiktive Validität von Einschätzungen bezüglich der Wirkungen von Bildungssoftware ist hingegen aufgrund fehlender Befunde zur Validität von Beurteilungsinstrumenten bzw. der unklaren Befundlage in der empirischen Forschung unbestimmt (z.B. Hasebrook, 1995; Schulmeister, 1996). Einer Gewichtung von Kriterien mangelt es an empirischen Belegen. Zimmer & Psaralidis (2000) stellen daher die prädiktive Validität von Kriterienkatalogen gänzlich in Frage. Sie stellen prinzipiell in Abrede, das Qualitätsevaluation von Bildungssoftware mittels Kriterienkatalog möglich ist. Sie gelangen aufgrund ihrer kritischen Betrachtung des vorherrschenden Evaluationsmodells der Wirkungsforschung zu dem Schluss, dass es generell keinen Sinn mache, von der Qualität einer Bildungssoftware zu sprechen, da nicht einzelne Merkmale der Software, sondern letztlich der mit ihrer Hilfe erzielte Lernerfolg deren Qualität bestimme. Dieser könne aber aufgrund der unbekannten Variablen, die auf das Lernergebnis Einfluss nehmen, nicht vorhergesagt werden, „denn die Qualität einer Lernsoftware wird erst in der Anwendungssituation selbst, also durch das Lernen mit ihr, also erst durch die Aktivitäten der Lernenden selbst, hergestellt. Daher kann ... gut verwendete, aber von Experten als schlecht bewertete Lernsoftware durchaus bessere Ergebnisse bei den Lernenden bringen als für gut befundene Lernsoftware. Weil die Qualität erst im Prozess des Lernens von den Lernenden selbst hergestellt wird, gegebenenfalls auch mit Unterstützung von Lehrenden, kann es keinen kausalen Zusammenhang zwischen objektiven Merkmalen und subjektiven Lernerfolgen geben" (S. 264/265). Die Argumentation von Zimmer & Psaralidis (2000) kann als konstruktivistische Betrachtung mediengestützten Lernens gewertet werden, nach der in letzter Konsequenz die Qualität konstruktivistischer Lernprozesse (und nicht die Qualität von Bildungssoftware) beurteilt werden muss, um Aussagen über das lernfördernde Potenzial einer gegebenen Bildungssoftware treffen zu können. Gestützt wird diese Auffassung durch Untersuchungen wie die von Reiser & Dick (1990). In dieser Untersuchung wurde eine von anerkannten Institutionen als hervorragend bewertete Bildungssoftware an Schülern erprobt. Es stellte sich heraus, dass die Lernerfolge der Schüler in keiner Weise der hervorragenden Bewertung entsprachen.

Untersuchungen wie diese machen differenzielle Methodeneffekte deutlich (Cronbach & Snow, 1977; Fricke, 2000). Hierunter versteht man, dass die Effekte einer Bildungsmaßnahme je nach Rahmenbedingungen (z.B. Lernin-

halt, Adressaten, curriculare Integration) unterschiedlich ausfallen können. So stellt Baumgartner (1995) mit Hinweis auf die Befunde von Fricke (1991) fest, dass empirische Untersuchungen mittels summativer Evaluationsverfahren gezeigt haben, „dass auch Lernprogramme, die mittels (didaktischer) Kriterienkataloge recht schlecht abschneiden, ..., in bestimmten Situationen durchaus erfolgreich und effektiv eingesetzt werden können" (S. 242). Entsprechende Befunde werden von Fricke (2000) mitgeteilt. Die Befunde zeigen, dass Ursachen der begrenzten prädiktiven Validität von Qualitätsbeurteilungen mittels Kriterienkatalogen auf eine mangelnde Berücksichtigung von Wechselwirkungen zwischen Softwaremerkmalen, Merkmalen der Inhaltsdomäne und kognitiven, emotionalen und volitionalen Lernvoraussetzungen auf Seiten der Lehrenden und Rahmenbedingungen des Softwareeinsatzes wesentlich mitbestimmt werden. Die empirischen Befunde stützen die Feststellung von Fricke, dass es „niemals 'die' Effektivität eines Lernprogramms geben kann" und „es auch nicht 'die' Validität eines einzelnen Programmmerkmals geben kann. „Diese werden je nach Rahmenbedingungen unterschiedliche Validitäten aufweisen, was zur Folge hat, dass ein Merkmal einer Lernsoftware in einem Fall ein Qualitätsmerkmal sein kann und in einem anderen Fall nicht" (Fricke, 2000, S. 81).

Starke Abhängigkeit der Urteilsvalidität von der Expertise des Beurteilers

Die Validität von Evaluationsurteilen ist in hohem Maße von der persönlichen Expertise der Beurteilenden abhängig. Die prädiktive Validität eines Kriterienkataloges kann letztlich nicht höher liegen als seine Reliabilität. Diese wiederum wird wesentlich von der Objektivität der Messungen, d.h. von der Übereinstimmung der Beurteilungen und der Expertise der Beurteilenden bestimmt. Expertise ist vor allem immer dann bedeutsam, wenn zum Beispiel Evaluationsurteile über die instruktionale Qualität bestimmter technischer Merkmale, vorgesehener Lernmöglichkeiten, zur Verfügung gestellter Hilfen abgegeben werden müssen und dabei Voraussetzungen auf Seiten der Lernenden und der Lernsituation und damit Wechselwirkungen zu berücksichtigen sind. Auf Seiten der Anwender von Kriterienkatalogen kann fachliche Expertise nicht immer als gegeben vorausgesetzt werden. Infolge unterschiedlichem individuellen Erfahrungshintergrund und unterschiedlichem Hintergrundwissen der Beurteilenden besteht auch bei sog. Experten nicht selten Unklarheit bezüglich der Bedeutung der Kriterien (Tergan, 2000 b).

Erläuterungen der Bedeutung von Kriterien als Anwendungshilfe innerhalb eines Beurteilungsinstruments schaffen hier eine gewisse Abhilfe. Zur Validitätserhöhung wäre ein intensives Beurteilertraining erforderlich, das auf die Entwicklung gemeinsamer Fachexpertise ausgerichtet ist. Hierbei geht es um den Erwerb theoretisch fundierten Wissens über die Bedeutung von Software-Merkmalen, Lernermerkmalen, didaktischen Methoden und deren Zusammenwirken im jeweils gegebenen situativen Kontext und dessen ganzheitlich-verstehensorientierter Anwendung bei der Beurteilung der pädagogischen Qualität von Bildungssoftware (vgl. McDougall & Squires, 1996; Schott, Krien, Sachse & Schubert, 2000; Tergan, 1998).

Geringe praktische Signifikanz der Qualitätskriterien

Auf das Problem der geringen praktischen Signifikanz von Qualitätskriterien verweisen Fricke (2000) und Zimmer & Psaralidis (2000). Fricke stellt hierzu u.a. fest: „Unabhängig vom Problem mangelnder Objektivitäts- und Zuverlässigkeitsmaße einzelner Merkmale wird man auf der Suche nach lernwirksamen Kriterien feststellen müssen, dass die Höhe der Korrelation zwischen Ausprägungsgrad der Qualitätskriterien und dem Lernerfolg in der Regel sehr gering ist. Der Korrelationskoeffizient mag statistisch signifikant ... sein, wegen seiner geringen Höhe weist er jedoch eine geringe praktische Signifikanz ... auf" (S. 77). Die praktische Bedeutsamkeit der Merkmale einer Bildungssoftware für die Vorhersage von Lernerfolg ist damit gering. Das Problem geringer praktischer Signifikanz der Merkmale von Bildungssoftware wurde bereits im Zusammenhang mit der Wirksamkeit von Medienmerkmalen von Clark & Craik (1992) diskutiert. Kerres (1998) geht auf dieselbe Problematik im Zusammenhang mit Erörterungen zum Design computerbasierter Lernumgebungen ein und verweist auf differenzielle Methodeneffekte (vgl. Fricke, 2000).

2.3 Anwendungsprobleme

Unhandlichkeit

Nach dem Wunsch von Anwendern sollten Kriterienkataloge „möglichst umfassend" (vgl. Glowalla, 1992, S. 40) und für die Evaluation unterschiedlicher Art von Bildungssoftware gleich gut verwendbar sein. Da sich Kriterienkataloge nur „schwer dem Verdacht der Unvollständigkeit entziehen" können (Baumgartner, 1995) und da mit der Entwicklung jedes neuen Katalogs neue Kriterien in die Diskussion einbezogen oder alte als zu unbestimmt definiert und weiter ausdifferenziert werden, können Unhandlichkeit und eine mangelnde Gerichtetheit der Kriterien für spezifische Verwendungszwecke die Folge sein. Dies kennzeichnet vor allem Evaluationsinstrumente der 1. Generation (vgl. Gräber, 1996).

Kriterienkataloge, die den Anspruch haben, über „vollständige" Kriterienlisten zu verfügen, eine sehr differenzierte Evaluation von Software anstreben, für unterschiedliche Bereiche und Intentionen der Evaluation, unterschiedliche Softwaretypen und unterschiedliche Zielgruppen geeignet sein sollen, umfassen zum Teil mehr als 300 Kriterien (Meier, 1995). Zu einem explosionsartigen Anwachsen der Kriterienliste kann es kommen, wenn immer wieder neue Kriterien einbezogen oder alte als zu unbestimmt definiert und noch weiter unterteilt werden (Baumgartner, 1995, S. 242). Dies ist dann der Fall, wenn nicht nur sog. eindimensionale Kriterien verwendet werden, sondern auch Wechselwirkungen mit anderen Kriterien sowie Rahmenbedingungen eines vorgesehenen Softwareeinsatzes mit berücksichtigt werden. Sofern keine „Filter" vorgesehen sind, um die Anzahl der für bestimmte Evaluationsaspekte bedeutsamen Fragen im Hinblick auf bestimmte Verwendungszwecke zu reduzieren und der Katalog nicht flexibel auf spezielle Ty-

pen von Bildungssoftware zugeschnitten werden kann, werden derartige Kriterienkataloge leicht unhandlich.

Mangelnde Gerichtetheit der Kriterien für spezifische Verwendungszwecke
Ein Anspruch an Kriterienkataloge ist in der Regel der, dass sie für die Evaluation unterschiedlicher Art von Bildungssoftware verwendbar sein sollten. Durch den Allgemeinheitsanspruch ergibt sich eine mangelnde Gerichtetheit der Kriterien für spezifische Verwendungszwecke. Squires & McDougall (1994) weisen darauf hin, dass bei der Anwendung von Kriterienkatalogen z.B. Unterschiede zwischen Gegenstandsbereichen hinsichtlich Strukturiertheit, Komplexität, Repräsentationsform und kognitiver Anforderung durch Anwendung gleicher Kriterien vernachlässigt werden (vgl. auch Jacobson & Spiro, 1994). In ähnlicher Weise werden Unterschiede zwischen Programmtypen nivelliert, deren Einsatz im Kontext unterschiedlicher Lehr-/Lernsituationen sinnvoll ist (z.B. Übungsprogramme, Tutorielle Systeme, Simulationen, Hypertexte/Hypermedien) (Mandl, Gruber & Renkl, 1992). Für hybride Formen von Bildungssoftware sowie didaktisch innovative Softwarelösungen sind Standardkataloge mit festgelegten Gewichtungen einzelner Kriterien ungeeignet (Heller, 1991). Um spezifischen Eigenschaften einer gegebenen Bildungssoftware und Rahmenbedingungen deren Anwendung im praktischen Kontext Rechnung zu tragen, bedarf es flexibel anwendbarer Instrumente.

Lösungsmöglichkeiten werden z.B. von Bangert-Drowns & Kozma (1989) und Jacobson & Spiro (1994) vorgestellt. So werden bei Bangert-Drowns & Kozma (1989) für unterschiedliche Systemtypen unterschiedliche Gewichtungen derselben Kriterien vorgenommen bzw. unterschiedliche Kriterien verwendet. In entsprechender Weise können Kriterien (z.B. mittels Filter) auf bestimmte Verwendungszwecke zugeschnitten werden. Jacobson & Spiro (1994) beschreiben in diesem Zusammenhang einen Weg zur Beurteilung der pädagogischen Qualität von Bildungssoftware, indem sie die jeweilige kognitive Anforderungssituation unter den Aspekten Strukturiertheit/Komplexität des Lerninhalts und Lernkompetenz der Studierenden präzisieren, unter der entsprechend ihrem theoretisch orientierten Ansatz mit unterschiedlichen Systemtypen effektiv gelernt wird. Sie sprechen dabei Empfehlungen aus, unter welchen Bedingungen die Anwendung eines bestimmten Typs von Bildungssoftware pädagogisch sinnvoll ist.

Mangelnde Anpassbarkeit von Katalogen an situative Gegebenheiten
Das Problem der mangelnden Anpassbarkeit vieler Kataloge hinsichtlich verwendeter Kriterien und zugrunde liegender Kriterien-Gewichtungen an vorliegende situative Gegebenheiten wird in der Kritik von Meier (2000) an dem auf Initiative des Arbeitskreises der Automobilindustrie entwickelten Kriterienkatalog AKAB deutlich. Meier (2000, S. 185) bemängelt, dass der „ganzheitliche Aspekt" einer zu beurteilenden Software bei inflexiblen Kriterienkatalogen „nicht genügend berücksichtigt" werden könne (vgl. Schott, 2000). „Da keine weiteren Kriterien aufgenommen bzw. konzeptionell nicht

relevante aussortiert werden können, kann der Beurteiler das Werkzeug nicht an innovative, aber auch an sog. 'Mischprogramme'... anpassen" (S. 185). Nach Meier wertet ein Kriterienkatalog, der wie AKAB diese Möglichkeit nicht bietet, nicht adäquat, „da bei der Endbewertung nicht berücksichtigte Kriterien auf '--'gesetzt werden" (S. 185).

Sowohl die mangelnde Gerichtetheit von Kriterien als auch die mangelnde Anpassbarkeit von Kriterienlisten an situative Gegebenheiten beeinflussen die Flexibilität der Nutzung von Kriterienkatalogen. Meier (2000) vermisst bei computergestützten Qualitätsbeurteilungsinstrumenten „die Integration eines Editors, der es ermöglicht, abweichende gute oder schlechte Aspekte von Lernsoftware in die Bewertung einzubeziehen" (S. 185). Hilfreich empfindet Meier eingebaute „Filter", die von vorn herein die Bewertung hinsichtlich Thema, Intention und Benutzern in Bahnen lenken. „Entsprechend nicht relevante Kriterien, die die Beurteilung verzerren könnten, könnten so zugunsten von entscheidenden Kriterien in der Bewertung 'umgangen' werden" (S. 185). Filterung erfordert jedoch eine à priori Festlegung relevanter Kriterien einschließlich deren Gewichtungen entsprechend einem zugrunde liegenden konzeptuellen Modell. Sofern die Filterung nach Kriterien pädagogischer Effektivität erfolgt, ist dabei eine Theorieorientierung unumgänglich (vgl. Baumgartner, 1995; Fricke, 1995, 2000; Schott, 2000). Squires & McDougall (1996) schlagen sog. generative Evaluationsinstrumente vor um den jeweils besonderen situativen Bedingungen der Softwareanwendung in der Beurteilung gerecht zu werden.

Informationsverlust durch zusammenfassende Bewertung

Um zu einem Gesamturteil bezüglich der Qualität einer Bildungssoftware zu gelangen, werden bei Kriterienkatalogen die Beurteilungen zu einzelnen Softwaremerkmalen in aller Regel zusammengefasst. In Fällen quantitativer Urteilsdaten bei Verwendung von Ratingskalen geschieht dies zum Beispiel durch Ermittlung eines Gesamtpunktwertes (u.a. Prichard et al., 1989). Bei qualitativen Beurteilungen werden die Einzelurteile zu einer zusammenfassenden evaluativen Stellungnahme aggregiert (vgl. Korbmacher zum Kriterienkatalog SODIS; Gräber, 1990 sowie Meier, 2000 zum Kriterienkatalog MEDA). Zusammenfassende Bewertungen sind bei der Qualitätsbeurteilung von Bildungssoftware hilfreich um z.B. auf der Basis des Gesamturteils eine Auswahl zwischen verschiedenen Softwareprodukten zu erleichtern. Gelegentlich werden in das Gesamturteil auch kurze Stellungnahmen bezüglich eigener Erfahrungen bei der Software-Anwendung integriert. Eine evaluative Stellungnahme ist beispielsweise bei Anwendung des SODIS-Beurteilungsinstruments erwünscht um Lehrern nützliche Hinweise für den Einsatz im Unterricht bereitzustellen (Korbmacher, 2000). Zusammenfassende Beurteilungen sind ferner sinnvoll um z.B. „Sinn und Nutzen eines Bildungsmediums bzw. -programms gegenüber Bildungsträgern, fördernden Institutionen, der Öffentlichkeit sowie potenziellen Anwendern zu begründen" (Tergan, 2000 a).

Bei der Datenaggregation entsteht jedoch zwangsläufig ein Informationsverlust. Informationsverlust bei der zusammenfassenden Bewertung kann dann zu Fehleinschätzungen bei der Qualitätsbeurteilung führen, wenn beispielsweise die Beurteilung einzelner prinzipiell bedeutsamer Kriterien für den Lernerfolg – z.B. Transparenz der Aufgabensituation, Vorliegen klarer Zielorientierungen – infolge einer Fokussierung hard- und softwaretechnischer Kriterien unberücksichtigt bleibt oder aber bedeutsame Kriterien und vergleichsweise unbedeutende Kriterien mit gleicher Gewichtung in das Gesamturteil eingehen. Im Gesamturteil wird nicht mehr erkennbar, auf welchen Kriterien dieses beruht und als wie bedeutsam die Kriterien bewertet wurden. Ein Vergleich der Beurteilungen verschiedener Beurteiler wird hierdurch erschwert.

Für die Qualitätsbeurteilung im Rahmen einer formativen Evaluation sind zusammenfassende Stellungnahmen weitgehend entbehrlich. Zum einen, weil die Beurteilung in dieser Phase eher Einzelaspekten der betrachteten Bildungssoftware gilt, zum anderen, weil sie im Falle einer angestrebten Software-Modifikation aufgrund fehlender Detailbeurteilung keine hinreichenden Hinweise für die Veränderung einzelner Softwareaspekte bieten.

Fehlende oder strittige Bewertungs- und Gewichtungsverfahren

Bei der Bewertung stellt sich die Frage nach dem zu verwendenden Bewertungs- und Gewichtungsverfahren und dessen theoretischer Begründung. Nach Baumgartner (1995) ist die Frage, wie die einzelnen Kriterien in Kriterienkatalogen zu gewichten sind, sehr umstritten. Er stellt fest: „Gerade die Gliederung und Gewichtung der einzelnen Kriterien ist entscheidend für eine vergleichende Bewertung und Auswahlentscheidung. In vielen Fällen wird daher die lerntheoretische Diskussion über Gewichtungsfragen ausgeklammert und den subjektiven Ansichten individueller AnwenderInnen, EvaluatorInnen, PädagogInnnen etc. überlassen". .... „Damit ist es auch mit der scheinbaren Objektivität von Kriterienkatalogen vorbei" (S. 242). Auch eine theorieorientierte qualitative Gewichtung nach dem Verfahren QWS (Qualitative Weight and Sum) wie sie von Baumgartner & Payr (1997) für die Auswahl von Bildungssoftware im Rahmen des EASA-Wettbewerbs vorgeschlagen wird, kann das Problem letztlich nicht lösen. Jede Beurteilung nicht eindeutig quantifizierbarer Merkmale steht und fällt mit der Expertise der Beurteilenden. Um die Urteilsobjektivität zu verbessern, wäre es zum einen wichtig, die jeweils zugrundeliegende theoretische Konzeption eines Kriterienkataloges offenzulegen. Es wäre zum anderen wichtig, theoriegeleitet umfangreiche Kenntnisse über den Einfluss von Softwaremerkmalen auf Lernergebnisse unter unterschiedlichen Rahmenbedingungen (z.B. Lernvoraussetzungen, Aufgabenstellungen) zusammenzutragen und an Beurteiler zur praktischen Anwendung bei der Nutzung von Kriterienlisten, z.B. durch ausführliche Kriterienbeschreibungen oder im Rahmen eines Beurteilertrainings weiter zu vermitteln (vgl. Schott, 2000). Empirische Befunde zeigen, dass Medienmerkmalen eine weitaus geringere Gewichtung zukommt, als dies in bestehenden Kriterienkatalogen durch Einbeziehung entsprechender Kriterien

berücksichtigt wird (Clark & Sugrue, 1990; Clark, 1994; Hasebrook, 1995; Schulmeister, 1996).

Fokussierung auf Stand-alone-Systeme

Die Qualitätsbeurteilung multimedialer Bildungssoftware mittels Kriterienkatalogen gilt üblicherweise ihrer Qualität als Stand-alone-Systeme für das Selbstlernen (vgl. Tergan, Hron & Mandl, 1992). Problematisch ist der Einsatz von Kriterienkatalogen dann, wenn Bildungssoftware integraler Bestandteil komplexer curricularer und technologiegestützter Lernumgebungen ist (z.B. Achtenhagen & John, 1992). Unter diesen Bedingungen kann zwar mittels Kriterienkatalog die technische Qualität einer Bildungssoftware beurteilt werden. Die Beurteilung der inhaltlichen und mediendidaktischen Qualität kann sich jedoch nicht mehr allein auf die Bildungssoftware selber beziehen, sondern muss zwangsläufig deren Integration in das gesamte Bildungsangebot betreffen, in dem die Software nur einen, nicht isoliert zu betrachtenden Teil darstellt. Dabei wäre die Angemessenheit der Software im Hinblick auf die besonderen situativen Bedingungen der Softwareanwendung, insbesondere kognitive, motivationale und volitionale Voraussetzungen auf Seiten der Lernenden und die didaktische Konzeption des Bildungsangebotes zu berücksichtigen. Für die Anwendung von Kriterienkatalogen hat dies Konsequenzen. Tergan (1998, S. 18) stellt mit Bezug auf Marchionini (1990) fest: „Advances in the applicability of checklists for software evaluation will depend on the inclusion of criteria addressing relevant features of integrated software and on the development of valid theories guiding a 'comprehensive and multi-faceted approach to evaluation'". Statt um Qualitätsbeurteilung von Bildungssoftware ginge es um die Beurteilung eines durch Bildungssoftware gestützten Lehr-Lern-Arrangements (vgl. Achtenhagen & John, 1992). Diese Beurteilung hat sinnvollerweise im Kontext eines umfassenden Qualitätssicherungsprozesses zu erfolgen. Sie kann ohne zusätzliche Einbeziehung empirischer Verfahren nicht mehr geleistet werden.

3. Fazit

Allgemein kann festgestellt werden, dass sich Kriterienkataloge für eine differenzierte Beschreibung der Merkmale von Bildungssoftware eignen. Für eine formative Evaluation im Rahmen der Qualitätssicherung bieten sie für Entwickler ein differenziertes Beschreibungsraster von Softwaremerkmalen, an denen sich Entwicklung orientieren kann. Dasselbe Beschreibungsraster eignet sich auch für einen Vergleich der Merkmale alternativer Varianten von Bildungssoftware und damit für die Unterstützung von Entscheidungsprozessen bei der Selektion, dem Kauf und der Anwendung von Bildungssoftware (vgl. SODIS: http://www.sodis.de/). Für eine Prädiktion der pädagogischen Effektivität und damit für eine Beantwortung der Frage „Kann mit einer bestimmten Lernsoftware effektiv gelernt werden", sind Kriterienkataloge hingegen wenig geeignet. Sie sind in der Regel zu starr und in-

flexibel, um den Besonderheiten einer Lernsoftware und situativen Bedingungen ihrer Anwendung gerecht zu werden. Kennzeichnend ist die Fokussierung von technischen und Design-Merkmalen und die Vernachlässigung von situativen Bedingungen der Softwarenutzung, d.h. individuellen Lernvoraussetzungen, Merkmalen des Lerngegenstandes, der Aufgabenstellung und der Lernsituation sowie Maßnahmen der Lernunterstützung und deren Wechselwirkung auf Lernprozesse. Kennzeichnend ist ferner, dass sich die Beurteilungskriterien in bestehenden Kriterienkatalogen vielfach an Paradigmen einer traditionellen behavioristischen bzw. kognitivistischen Form der Lernförderung orientieren, bei denen die Optimierung von Softwaremerkmalen und Vermittlungsformen im Fokus steht. Einer konstruktivistischen, auf den Lernprozess orientierten Sichtweise wird jedoch kaum Rechnung getragen (Squires & McDougall, 1996).

Die Anwendung eines Kriterienkatalogs zur Qualitätsbeurteilung kann zwar wichtige Hinweise zur technischen, inhaltlichen und didaktischen Qualität einer Bildungssoftware liefern, die von Experten vor dem Hintergrund eines bestimmten State-of-the-Art des Software-Designs für eine erste Qualitätseinschätzung verwendet werden können. Für eine Beurteilung der Effektivität von Bildungssoftware ist jedoch die Hinzuziehung empirischer Methoden unabdingbar.

4. Perspektiven

Derzeit werden unterschiedliche Ansätze für die Nutzung von Kriterienkatalogen und Checklisten im Rahmen der Qualitätsevaluation von Bildungssoftware präferiert. Diese betreffen zum einen Maßnahmen zur Verbesserung der prädiktiven Validität von Evaluationsurteilen, zum anderen Modelle der gezielten Verwendung der Evaluationsinstrumente im Rahmen eines mehrstufigen und umfassenden Ansatzes der Qualitätsbeurteilung.

Vorschläge für Maßnahmen zur Verbesserung der prädiktiven Validität gelten einer stärkeren Orientierung der Qualitätsbeurteilung an Theorien des Instruktionsdesigns (Baumgartner, 1995; Fricke, 1995; Schott, 2000) sowie Bestrebungen, die in Lernprozessen bedeutsamen Wechselwirkungen von Softwaremerkmalen, Adressatenmerkmalen (z.B. Vorwissen, computer literacy), Aufgabenmerkmalen und situativen Rahmenbedingungen (vgl. Bangert-Drowns & Kozma, 1989; Tergan, 1998) sowie den Implikationen des konstruktivistischen Lernparadigmas für die Qualitätsbeurteilung von Bildungssoftware zu entsprechen (Squires & McDougall, 1996).

Die Orientierung an Instruktionsdesign-Ansätzen erscheint in diesem Zusammenhang als ein erster Schritt (Baumgartner, 1995; Fricke, 2000). Die Merkmale von Bildungssoftware sowie die Bedingungen, unter denen diese eingesetzt wird, sind jedoch zu unterschiedlich, differenzielle Methodeneffekte zu wahrscheinlich, so dass allgemeine Instruktionsdesign-Modelle sowie einzelne empirische Befunde nur einen Theorierahmen abgeben können, vor dessen Hintergrund jeweils vom Beurteilenden selber situationsspe-

zifische Annahmen und Vorhersagemodelle zu entwickeln und letztlich auch zu überprüfen sind (Fricke, 2000). Bestehende Kriterienkataloge sind durch ein vergleichsweise starres Beurteilungsraster und einen fehlenden Zuschnitt auf die besonderen Bedingungen einer Bildungssoftware und deren Einsatz im Bildungskontext gekennzeichnet. Sie sollten daher vor ihrer Verwendung danach bewertet werden, ob implizite Annahmen zur pädagogischen Qualität von Bildungssoftware jenen Annahmen entsprechen, die vom Beurteiler selber unter den gegebenen Bedingungen eines Softwareeinsatzes als angemessen erachtet werden sowie danach, ob die besonderen Bedingungen des Softwareeinsatzes ein flexibleres Instrumentatium erforderlich machen.

Sinnvoll für eine Verbesserung der prädiktiven Validität von Bildungssoftware erscheint eine an Kriterien orientierte, jedoch auf wissenschaftlicher Expertise gründende verstehensorientierte, ganzheitliche Qualitätsbeurteilung (Squires & McDougall, 1996; Tergan, 1998; Schott, 2000). Diese berücksichtigt stärker Wechselwirkungsbeziehungen von Eigenschaften der Lernenden, der Lerninhalte, der Lehrpersonen, Lehrmethoden und Medien, die speziellen situativen Rahmenbedingungen, unter denen die Bildungssoftware eingesetzt werden soll, Informationen aus Praxisbeispielen eines erfolgreichen Softwareeinsatzes sowie sogenannte hochinferente Variablen (Fricke, 2000), deren Auswirkungen auf den Lernerfolg empirisch nachgewiesen werden konnten. Entsprechende Konzeptionen für die Software-Evaluation wurden u.a. von Tergan, Hron & Mandl (1992), Squires & McDougal (1996) sowie Tergan (1998) vorgestellt.

Durch die Einbeziehung anwendungsorientierter Kriterien werden dabei die Rahmenbedingungen eines Softwareeinsatzes stärker in den Vordergrund gerückt. Ein derartiges Vorgehen wird beispielsweise im Rahmen des Bewertungsprozesses zur Verleihung des Mediendidaktischen Hochschulpreises (Medida-Prix) praktiziert. Zu diesen Kriterien gehören z.B. Kriterien wie „Didaktische Motivation", „Nachhaltigkeit", „Übertragbarkeit", „Alltagstauglichkeit" und „Qualitätssicherung" einer Bildungssoftware. Didaktische Motivation liegt dann vor, wenn der didaktische Nutzen im Sinne eines Mehr-Wertes deutlich wird. Nachhaltigkeit wird dann attestiert, wenn nachgewiesen werden kann, dass eine curriculare Integration in eine umfassendere Bildungsmaßnahme sowohl konzeptuell, finanziell als auch personell sichergestellt ist. Übertragbarkeit betrifft die Frage der technischen, didaktischen und organisatorischen Portierbarkeit eines Projekts in andere Fachbereiche. Mediale Innovation wird dann attestiert, wenn ein (Zusatz-)nutzen der verwendeten Medien deutlich gemacht werden kann. Alltagstauglichkeit bezieht sich u.a. auf Möglichkeiten der Adaptierung und Anpassung eines Bildungsangebotes an sich verändernde situative Bedingungen. Qualitätssicherung wird als Auswahlkriterium verwendet, um grundlegende Standards der Qualität von Bildungsangeboten sicher zu stellen.

Da eine prädiktive Evaluation der Effektivität von Bildungssoftware allein auf der Basis einer Expertenbeurteilung mittels Kriterienkatalog nicht leist-

bar ist, empfiehlt sich eine mehrstufige und kombinierte Anwendung von Expertenbeurteilungs- und empirischen Evaluationsverfahren. Bei einer kombinierten Anwendung von Expertenbeurteilungs- und empirischen Evaluationsverfahren sollte nach Meier Bildungssoftware, die eine Vorprüfung unter Verwendung eines Kriterienkataloges erfolgreich bestanden hat, „anschließend durch Benutzer der entsprechenden Zielgruppe getestet werden" (Meier, 2000, S. 188). Hierbei stehen sowohl der Kriterienkatalog selber als auch das zugrunde liegende Vorhersagemodell auf dem Prüfstand. „The results must be evaluated in the light of the model and, if necessary, must be the basis for the revision of models that have proven to be ineffective" (Tergan et al., 1992, S. 154). Eine entsprechende Vorgehensweise ist beispielsweise bei der Volkswagen-AG Standard (Meier, 1995). Informelle empirische Wirkungsanalysen im Rahmen von Softwareerprobungen unter Einbeziehung einzelner Testpersonen der anvisierten Zielgruppe einer Bildungssoftware unter möglichst realitätsnahen Anwendungsbedingungen machen zwar zusätzlichen Aufwand. Sie können jedoch maßgeblich zur Verbesserung der Urteilsbildung beitragen und insbesondere Aufschluss darüber geben, ob nicht nur die einfache Übertragung von Faktenwissen, Routinefertigkeiten und Regelwissen unter den realen Bedingungen der Anwendungssituation gelingt, sondern auch die „Re-Konstruktion von Wissen und Kompetenzen" (Baumgartner, 1995; Tergan, 1998, 2000 c; Freibichler, 2000; Reinmann-Rothmeier & Mandl, 2000).

Eine umfassende Qualitätsbeurteilung von bereits existierender Bildungssoftware kann sich nicht allein auf Produktmerkmale beschränken, sondern hat im Rahmen eines maßgeschneiderten Vorgehens (Mandl & Reinmann-Rothmeier, 2000) zentrale Aspekte seiner Planung, Entwicklung und Bedingungen seiner Anwendung zu berücksichtigen. Sie hat auf unterschiedlichen Ebenen mit unterschiedlichen Perspektiven und Zielsetzungen und unter Einsatz unterschiedlicher Methoden zu erfolgen (vgl. Schenkel, 2000). Das Evaluationssystem EVAluationsnetz (http://www.evaluationsnetz.com) bietet hier für die Zielgruppe der Praktiker (v.a. Softwareanwender) Anleitungen und Hintergrundinformationen zur Qualitätsevaluation. Für eine Beurteilung der Wirkungen einer Bildungssoftware kommt einer Erprobung an potenziellen Nutzern der Software unter möglichst praxisnahen Bedingungen eine zentrale Rolle zu. Fachexpertise der Beurteiler kann empirische Erprobung nicht ersetzen. Mittels Kriterienkatalogen kann zwar von Experten die Qualität des Designs beurteilt werden. Sofern die Kriterien nicht nur technische, sondern vor allem didaktische, nutzerbezogene und situative Aspekte der Softwarenutzung betreffen und das Beurteilungsinstrument auf die besonderen Bedingungen der Softwarenutzung anpassbar ist, können auf der Grundlage der Beurteilung begründete Schlüsse bezüglich der grundsätzlichen Eignung einer Bildungssoftware zur Unterstützung von Lernprozessen gezogen werden. Nur eine Erprobung unter realitätsnahen Bedingungen kann aber valide Hinweise auf die pädagogische Effektivität liefern. Der Anwendbarkeit von Kriterienkatalogen werden hierdurch enge Grenzen gesetzt. Innerhalb dieser Grenzen erweist sich jedoch eine von geeigneten Fach-

experten mittels geeignetem Kriterienkatalog durchgeführte ganzheitlich orientierte Qualitätsbeurteilung von Bildungssoftware ohne Alternative.

Literatur

Achtenhagen, F. & John, E.G. (Hrsg.). (1992). Mehrdimensionale Lehr-Lern-Arrangements: Innovationen in der kaufmännischen Aus- und Weiterbildung. Wiesbaden: Gabler.

Astleitner, H. (1997). Lernen in Informationsnetzen. Frankfurt: Lang.

Bangert-Drowns, R.L. & Kozma, R.B. (1989). Assessing the design of instructional software. Journal of Research on Computing in Education, 3(21), 241-262.

Baumgartner, P. (1995). Didaktische Anforderungen an (multimediale) Bildungssoftware. In L.J. Issing & P. Klimsa (Hrsg.). Information und Lernen mit Multimedia (S. 241-252). Weinheim: Psychologie Verlags Union.

Baumgartner, P. & Payr, S. (1997). Methods and practice of software evaluation. The case of the European Academic Software Award. Proceedings of the ED-Media '97 World Conference on Educational Multimedia and Hypermedia (44-50). Charlottesville: ACCE.

Brown, J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-42.

Clark, R. (1994) Media will never influence learning. Educational Technology Research and Development, 42 (2), 21-29.

Clark, R.E. & Craik, T.G. (1992). Research and theory on multimedia-learning effects. In M. Giardina (Ed.), Interactive multimedia learning environments. Human factors and technical considerations on design issues (NATO ASI Series. Series F: Computer and Systems Sciences, Vol. 93 (pp. 19-30). Berlin/Heidelberg: Springer.

Clark, R.E. & Sugrue, B.M. (1990). North American disputes about research on learning from media. International Journal of Educational Research, 14 (6), 507-519.

Cronbach, L.J. & Snow, R.E. (1977). Aptitudes and instructional methods. New York: Wiley.

Doll, C.A. (1987). Evaluating educational software. Chicago/London: Americal Library Association.

Fezzardi, G., Hasebrook, J. & Glowalla, U. (1992). MEM - Ein Hypermediasystem zur Entwicklung, Evaluation und Durchführung computerunterstützter Aus- und Weiterbildung. Handbuch, Gießen.

Freibichler, H. (2000). Protokolle von Lernprozessen. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.). Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 304-328). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Fricke, R. (1991). Zur Effektivität computer- und videounterstützter Lernprogramme. In R.S. Jäger u.a. (Hrsg.). Computerunterstütztes Lernen. Beiheft 2 zur Zeitschrift Empirische Pädagogik, 167-204.

Fricke, R. (1995). Evaluation von Multimedia. In L.J. Issing & P. Klimsa (1995). Information und Lernen mit Multimedia (401-413). Weinheim: Psychologie Verlags Union.

Fricke, R. (2000). Qualitätsbeurteilung durch Kriterienkataloge. Auf der Suche nach validen Vorhersagemodellen. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme.

Evaluationsmethoden auf dem Prüfstand (S. 75-88). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Friedrich, H.F. & Hron, A. (in Druck). Überlegungen zur Gestaltung und Evaluation virtueller Seminare. In H.M. Niegemann & K. Treumann (Hrsg.), Lehren und Lernen mit neuen Medien. Münster: Waxmann

Glowalla, U. (1992). Evaluation computerunterstützten Lernens. In U. Glowalla & E. Schoop (Hrsg.), Hypertext und Multimedia: Neue Wege in der computerunterstützten Aus- und Weiterbildung, S. 39-40. Berlin/Heidelberg: Springer-Verlag.

Gräber, W. (1990). Das Instrument MEDA. Ein Verfahren zur Beschreibung und Bewertung von Lernprogrammen. Institut für die Pädagogik der Naturwissenschaften (IPN), Kiel.

Gräber, W. (1996). Kriterien und Verfahren zur Sicherung der Qualität von Lernsoftware in der beruflichen Weiterbildung. Kiel: Institut für die Pädagogik der Naturwissenschaften.

Hasebrook, J.P. (1995). Lernen mit Multimedia. Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 9(2), 95-103.

Heller, R. (1991). Evaluating software: a review of the options. Computers and Education, 17(4), 285-291.

Ingenkamp, K. (1995). Lehrbuch der Pädagogischen Diagnostik. Weinheim: Beltz.

Jacobs, G. (1998). Evaluating courseware: some critical questions. Innovations in Education and Training International, 35(1), 3-8.

Jacobson, M.J. &Spiro, R.J. (1994) A framework for the contextual analysis of technology-based learning environments. Journal of Computing in Higher Education, 2, 5, 3-32.

Jolicoer, K. & Berger, D. (1986). Do we really know what makes educational software effective? A call for empirical research. Educational Technology, 6(25), 7-11.

Kerres, M. (1998). Multimediale und telemediale Lernumgebungen. Konzeption und Entwicklung. München: Oldenburg Verlag.

Korbmacher, K. (2000). Evaluation von Bildungssoftware auf der Basis von SODIS. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 190-216). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Lottmann, A. (2000). Die multimediale Bildungssoftware „Informations- und Kommunikationstechniken im Handwerk – IKTH". Ein Modellversuch zur beruflichen Bildung. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 126-136). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Mandl, H., Gruber, H. & Renkl, A. (1992), Lernen mit dem Computer. Empirisch-pädagogische Forschung in der BRD zwischen 1970 und 1990. Forschungsbericht 7. Ludwig Maximilians Universität München. Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie.

Mandl, H. & Reinmann-Rothmeier, G. (2000). Vom Qualitätsbewusstsein über Selbstevaluation und maßgeschneidertes Vorgehen zur Transfersicherung. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 89-105). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Marchionini, G. (1990). Evaluating hypermedia-based learning. In D.H. Jonassen & H. Mandl (Eds.), Designing hypermedia for learning. NATO ASI Series. Series F: Computer and System sciences. Vol. 67 (255-373).

Meier, A. (1995) Qualitätsbeurteilung von Bildungssoftware durch Kriterienkataloge. In P. Schenkel and H. Holz (Hrsg.), Evaluation multimedialer Lernprogramme und Lernkonzepte (pp. 149-191). Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen Verlag und Software GmbH.

Meier, A. (2000). MEDA und AKAB: Zwei Kriterienkataloge auf dem Prüfstand. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 164-189). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Prichard, W.H. Jr., Micceri, Th. & Barrett, A.J. (1989). A review of computer-based training materials: Current state of the art (instruction and interaction). Educational Technology, 16-22, July 1989.

Reiser, R.A. & Dick, W. (1990). Evaluation of instructional software. Educational Technology. Research and Development, 38(3), 43-50.

Schenkel, P., Tergan, S.-O. & Lottmann, A. (2000) (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand. Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Schenkel,P. (2000). Ebenen und Prozesse der Evaluation. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 52-74). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Schott, F. (2000). Evaluation aus theoriegeleiteter, ganzheitlicher Sicht. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 106-124). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Schott, F., Krien, F., Sachse, S. & Schubert, T. (2000). Evaluation von multimedialer Bildungssoftware auf der Basis von ELISE (1.0). Ein Ansatz zu einer theorie-, adressaten- und anwendungsorientierten Methode zur Evaluation von multimedialen Lern- und Evaluationssystemen. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 217-242). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Schulmeister, R. (1996). Grundlagen hypermedialer Lernsysteme. Theorie, Didaktik, Design. Bonn/Paris: Addison-Wesley.

Squires, D. & McDougall, A. (1994). Choosing and using educational software: a teachers´ guide. Falmer Press, London.

Squires, D. & McDougall, A. (1996). Software evaluation: a situated approach. Journal of Computer Assisted Learning, 12, 164-161.

Tergan, S.-O. (1998). Checklists for the evaluation of educational software: critical review and prospects. Innovations in Education and Training International, 35(1), 9-20.

Tergan, S.-O. (2000 a). Grundlagen der Evaluation. Ein Überblick. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern-und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 22-51). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Tergan, S.-O. (2000 b). Bildungssoftware im Urteil von Experten. 10 + 1 Leitfragen zur Evaluation. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 137-163). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Tergan, S.-O. (2000 c). Vergleichende Bewertung von Methoden zur Beurteilung der Qualität von Lern- und Informationssystemen. Fazit eines Methodenvergleichs. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 329-347). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Tergan, S.-O., Hron, A. & Mandl, H. (1992). Computer-based systems for open learning. In S.-O. Tergan, J.J. Sparkes, C. Hitchcock, A.R. Keye, A. Hron & H. Mandl (G. Zimmer & D. Blume, Eds.), Open learning and distance education with computer support (pp. 97-195). Band 4 der Reihe „Multimediales Lernen in der Berufsbildung". Nürnberg: BW Bildung und Wissen Verlag und Software GmbH.

Tolhurst, D. (1992). A checklist for evaluating content-based hypertext computer software. Educational Technology, 32(3), 17-21.

Willumeit, H., Gediga, G. & Hamborg, K.-C. (1996). IsoMetricsL. Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10. Ergonomie & Informatik. Der ISO 9241-Evaluator. Mitteilungen des Fachausschusses 2.3 „Ergonomie in der Informatik". März 1996 (5-12).

Zimmer, G. & Psaralidis, E. (2000). „Der Lernerfolg bestimmt die Qualität einer Bildungssoftware!" Evaluation von Lernerfolg als logische Rekonstruktion von Handlungen. In P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.), Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand (S. 22-51). Reihe Multimediales Lernen in der Berufsbildung. Nürnberg: BW Bildung und Wissen.

Anschrift des Autors:
Dr. Sigmar-Olaf Tergan
Institut für Wissensmedien (IWM)
Abteilung Angewandte Kognitionswissenschaft
Konrad-Adenauer-Str. 40, 72072 Tübingen, Germany
Tel.: + 49-7071/979-227, Fax: + 49-7071/979-100
E-mail: s.tergan@iwm-kmrc.de