

2 Theoretische Herleitung

Die theoretische Basis für die Beantwortung der Forschungsfrage stellen aktuelle Erkenntnisse aus der pädagogischen Diagnostik, der Kognitionspsychologie, und insbesondere der *Cognitive Load Theory* (CLT), sowie der Usability-Forschung dar. In den folgenden Kapiteln wird die Forschungsfrage aus den eben genannten theoretischen Zugängen hergeleitet.

2.1 Diagnostische Grundlagen

Die Online-Diagnostik *otulea* soll sowohl für genuin diagnostische Tätigkeiten als auch für Tätigkeiten im Kontext förderdiagnostischer Absichten entwickelt werden. In den folgenden zwei Kapiteln werden daher grundlegende Erkenntnisse zur Pädagogischen Diagnostik sowie Prinzipien der Förderdiagnostik erläutert.

2.1.1 Pädagogische Diagnostik

Die Pädagogische Diagnostik umfasst Tätigkeiten, die bei den Lernenden die Voraussetzungen und Bedingungen planmäßiger Lehr- und Lernprozesse ermitteln, Lernprozesse analysieren sowie Lernergebnisse feststellen (Ingenkamp & Lissmann, 2008, S. 13). Die Ergebnisse der Pädagogischen Diagnostik dienen der Zuweisung zu Lerngruppen oder individuellen Förderprogrammen, um beispielsweise durch weniger heterogene Gruppen ein "effizienteres" Lernen und Fördern zu ermöglichen. Übergeordnetes Ziel der Pädagogischen Diagnostik ist somit die Optimierung individuellen Lernens. Zudem dient sie der Steuerung des Bildungsnachwuchses, z. B. durch die differenzierte Abbildung des Kompetenzstands und anschließender Empfehlungen für den weiteren Bildungsweg oder der Erteilung von Qualifikationen, beispielsweise mittels der Bestätigung eines bestimmten Kompetenzniveaus.

Zunehmend wurden und werden unter den Funktionen der pädagogischen Diagnostik auch die Förderung des Lernens, die Verbesserung des Unterrichts, die Schüler-/Schülerinnenberatung und die Platzierung im Bildungssystem genannt (Lindquist, 1950; zitiert nach Ingenkamp & Lissmann, 2008, S. 22). Weiterhin werden als Ziele des Einsatzes von diagnostischen Verfahren die Diagnose von Stärken und Schwächen des Lernens, die Planung und Verbesserung des Unterrichts sowie die Evaluation von Leistung genannt (Phye, 1996; zitiert nach Ingenkamp & Lissmann, 2008). Die Pädagogische Diagnostik kann also auch einen Bestandteil einer Evaluation darstellen, doch bezieht sich der Begriff Evaluation primär auf die Bewertung von Interventionen (Wottawa, 2006, S. 650),

beispielsweise Kursen oder Kursentwicklungen oder auch auf den Prozess, wie sich die Bewertungen und das Bewertungsschema entwickelt haben (Taras, 2005, S. 467) und weniger auf die Bewertung personenbezogener Merkmale.

Bei der Pädagogischen Diagnostik handelt es sich in ihrer Konsequenz meist um selektive Tätigkeiten und somit um selektive Diagnostik. In der selektiven Diagnostik, als spezielle Ausprägung der pädagogischen Diagnostik, wird eine soziale oder sachliche Bezugsnorm herangezogen, um entweder zu prüfen, wie sich die Performanz einer Person im Vergleich zu einer anderen Person bzw. Gruppe oder zu einem fachlich definierten Anspruch verhält. Zudem wird die selektive Diagnostik von einer außen stehenden Instanz (z. B. Lehrperson) durchgeführt, bei der die Ergebnisse verbleiben (Schule, Arbeitgeber, Arbeitsagentur etc.). Als Konsequenz der selektiven Diagnostik wird – wie der Name ausdrückt – selektiert, d. h. es wird eine soziale oder sachliche und keine individuelle Bezugsnorm für die Leistungsbewertung herangezogen und eine Person wird aufgrund ihrer Performanz einer bestimmten Gruppe oder auch Institution zugeordnet. Die/Der Lernende steht dabei nicht im Fokus, was wiederum einen negativen Effekt auf die Lernmotivation haben kann (vgl. z. B. Rheinberg & Krug, 2005). Historisch betrachtet gewann die pädagogische Diagnostik zur Erteilung von Qualifikationen (wie z. B. die Erteilung von Zugangsberechtigung zur Ausbildung/zum Studium) in den letzten zwei Jahrhunderten zunehmend an Bedeutung⁴. Mit der zunehmenden Demokratisierung und der Ausschaltung von Geburtsrechten und Ämterkauf wurden individuelle Lernerfolge zunehmend wichtiger. Allerdings verlief eine Diagnostik bis dahin eher intuitiv und nicht mit Hilfe wissenschaftlicher Methoden. Ein vielfältiges Instrumentarium möglichst objektiver Verfahren ist in den vergangenen hundert Jahren entstanden (Ingenkamp & Lissmann, 2008, S. 22), mit Hilfe derer (unter Berücksichtigung der Gütekriterien) Beobachtungs- und Befragungsergebnisse ausgewertet werden, um die aktuelle und/oder zukünftige Performanz zu beschreiben. Ein Beispiel, das bis in die Gegenwart hineinreicht, ist der Einsatz von Schulleistungstest. Von der Pädagogischen Diagnostik abzugrenzen ist die pädagogische Forschung. In der pädagogischen Forschung wird diagnostiziert, um allgemeine Erkenntnisse über beispielsweise die Angemessenheit bestimmter didaktischer Vorgehensweisen oder den Mediengebrauch zu gewinnen. Auch wenn hier diagnostische Methoden in der Pädagogik genutzt werden, sind diese Erhebungen nicht der pädagogischen Diagnostik zuzuordnen. Die pädagogische Diagnostik zeichnet sich somit – im Unterschied zur pädagogischen Forschung – durch den Bezug auf aktuelle Maßnahmen bei einzelnen Lernenden oder einer Gruppe von Lernenden aus (Ingenkamp & Lissmann, 2008, S. 19). Im internati-

⁴ Wobei anzumerken ist, dass auch bereits zu Beginn des 7. Jahrhunderts in China Beamtenprüfungen durchgeführt wurden (Glöckner, 2013, S. 194).

onalen Raum ist für den Begriff der Diagnostik im Allgemeinen der Begriff *Assessment* (engl. für *Bewertung*) gebräuchlich. Der Begriff *Diagnostic Assessment* ist am ehesten mit pädagogischer Diagnostik zu übersetzen. *Diagnostic Assessment* wird durchgeführt, um Vorwissen von Lernenden zu erheben und um zukünftiges Lernen zu optimieren (Crisp, 2007, S. 254). Um explizit auf einen pädagogischen Bezug hinzuweisen wird auch der Begriff *Educational Assessment* gewählt, jedoch wird von der Spezifizierung „*Educational*“ selten Gebrauch gemacht⁵. Weniger im deutschen, jedoch im englischen Sprachraum werden differenzierend, u.a. im pädagogischen Kontext, die Begriffe *Large-Scale Assessment* und *Summative Assessment* genutzt. *Large-Scale Assessment* sind Schulleistungsuntersuchungen mit einer großen Anzahl von Teilnehmenden.

„Large-scale educational assessment consists of those tests administered to sizable numbers of people for such purposes as placement, course credit, graduation, educational admissions, and school accountability. It includes group-administered, standardizes tests“ (Bennett, 1998, S. 7).

Large-Scale Assessments zeichnen sich demnach durch ähnliche Ziele wie die Pädagogische Diagnostik aus, allerdings mit der Bedingung große Stichproben einzusetzen. Charakteristisch ist dabei die Durchführung von vorwiegend politischen Entscheidungsträgern und Bildungspolitikern bzw. führenden Bildungsinstitutionen, um entweder Bildungsprogramme zu evaluieren oder um Lernziele bei Schülern und Schülerinnen festzustellen (Pellegrino, Chudowsky & Glaser, 2001). Die Datenhoheit liegt dabei bei der forschenden Instanz. Ein bekanntes Beispiel ist die internationale Vergleichsstudie PISA (*Programme for International Student Assessment*).

Beim *Summative Assessment* handelt es sich in seiner Konsequenz meist um eine selektive Diagnostik.

„Summative assessment is intended to summarize student accomplishment by making judgment or determining a grade“ (Stödsberg, 2012, S. 595).

Summative Assessment bedeutet, Hinweise und Ergebnisse hinsichtlich eines vorgegebenen Standards auszuwerten. Dieser Punkt hat in Bezug auf *Assessment* den Charakter der Endgültigkeit. Der Prozesscharakter wird beim *Summative*

⁵ So findet sich z. B. im Glossar des E-Assessment Handbooks (Crisp, 2007) zwar eine Definition des Begriffs "Diagnostic Assessment", jedoch nicht für "Educational Assessment".

Assessment nicht berücksichtigt (Taras, 2005, S. 468). Sowohl das *Large-Scale* als auch das *Summative Assessment* haben eine soziale Bezugsnorm (im Gegensatz zum *Formative Assessment*, welches auf eine individuelle Bezugsnorm referenziert und nicht die einmalige Bewertung einer Person zum Zwecke der Selektion zum Ziel hat (vgl. nachfolgendes Kap. 2.1.2) und erheben den Anspruch der Statuserhebung.

Der deutsche Begriff Pädagogische Diagnostik scheint somit eine eindeutige Domänenspezifität (nämlich Pädagogik) vorzuweisen. Im Englischen werden mit dem Begriff *Assessment* nicht nur pädagogische Fragestellungen assoziiert. Die Differenzierungen in *Summative* und *Large-Scale Assessment* weisen hingegen eine Spezifität auf, zu denen sich im Deutschen keine äquivalenten Begriffe finden lassen. Die englischen Begriffe wurden in den deutschen Sprachraum übernommen, wobei der englische Begriff *summative* in „summativ“ übersetzt wurde, für den englischen Begriff *Large-Scale* besteht allerdings keine äquivalente deutsche Übersetzung. Dieser Abgrenzungsansatz zwischen deutschen und englischen Begriffen sowie deren Ausprägungen macht deutlich, dass Begriffe aus dem Englischen verwendet werden, deren Verwendung hingegen im Deutschen nicht immer äquivalent stattfindet. Um ein einheitliches und präzises Verständnis der Begrifflichkeiten zu ermöglichen, werden in den folgenden Kapiteln vorzugsweise die deutschen Begriffe verwendet. In einigen Ausnahmefällen wird auf englische Begriffe zurückgegriffen, da es sich entweder um im deutschen Sprachraum etablierte Anglizismen handelt und/oder der englische Begriff ein präziseres Verständnis ermöglicht.

Zusammenfassend liefert die pädagogische Diagnostik zwar Ergebnisse zum Leistungsstand einer Person, allerdings nicht (bzw. nur sehr bedingt) hinsichtlich abzuleitender individueller Fördermaßnahmen. Faktoren, die außerhalb der Prüfungsleistung liegen, wie z. B. Lernhemmungen, -behinderungen, das soziale Umfeld oder fehlende Passungen zwischen Unterrichtsmethode und Test, werden ebenfalls nicht berücksichtigt. Diese Abgrenzung der Funktionen und Ziele pädagogischer Diagnostik macht deutlich, dass der Bereich der pädagogischen Diagnostik auch das Thema Förderung umfasst, allerdings ohne explizit der Förderdiagnostik gerecht zu werden (vgl. folgendes Kap. 2.1.2). Diese fehlenden Faktoren werden in der Pädagogischen Förderdiagnostik einbezogen (Engel, 2008, S. 33). Die Pädagogische Diagnostik ist damit keineswegs als Gegenpool zur Förderdiagnostik zu verstehen, vielmehr stellt Pädagogische Diagnostik eine Erweiterung zur pädagogischen und speziell zur selektiven Diagnostik sowie eine Spezifizierung (u. a. in der Darstellung der individuellen Performanz) dar.

2.1.2 Förderdiagnostik

Die pädagogische Förderdiagnostik geht über reine diagnostische Tätigkeiten hinaus, um nach Ansatzpunkten für eine Förderung sowie nach veränderbaren Bedingungen in der Lernsituation zu suchen (Dupuis & Kerkhoff, 1992, S. 212). Pädagogische Förderung besteht in der Bereitstellung und Durchführung individuell zugeschnittener Angebote, wenn die (schulischen) Standardangebote nicht ausreichend für die Entwicklung der Person sind (Kretschmann, 2006, S. 140). Die in der pädagogischen Diagnostik gewonnenen Informationen werden genutzt, um individuelle Fördermaßnahmen abzuleiten (Breitenbach, 2007, S. 40), wobei die Daten bei der diagnostizierenden Instanz verbleiben. Sechs Prinzipien sind in der Förderdiagnostik vorherrschend (in Anlehnung an Rittmeyer, 2005; Schönrade & Pütz, 2004; vgl. Wolf, Koppel & Schwedes, 2011):

1. Individualität
2. Prozessorientierung
3. Wechselverhältnis von Diagnose und Intervention
4. Einbettung in das reale Umfeld der Teilnehmenden
5. Berücksichtigung von Stärken und Schwächen
6. Anwendung von Beobachtungsverfahren und Fehleranalysen

Im Folgenden werden die Prinzipien näher erläutert:

1. Individualität: Wie bereits vorangegangen angedeutet ist die Förderdiagnostik am individuellen Lernenden orientiert. Ziel ist, Lernschwierigkeiten und deren Entstehung zu ergünden. Bei Bedarf werden Veränderungen von Lernprozessen initiiert. Der Förderbedarf wird deskriptiv erfasst und es werden Hypothesen über mögliche Ursachen für Lernschwierigkeiten gebildet, um mögliche Interventionsstrategien zu entwickeln (Engel, 2008, S. 34).

2. Prozessorientierung: Im Kontext förderdiagnostischen Handelns wird davon ausgegangen, dass Entwicklungs- und Lernfähigkeit des Lernenden dynamisch veränderbar sind. Daher ist die Förderdiagnostik immer über einen längeren Zeitraum durchzuführen. In diesem werden mehrmals Lernstandserhebungen vorgenommen und Fördermaßnahmen abgeleitet (Petermann & Petermann, 2006, S. 2). Folglich werden die Begriffe Lernprozessdiagnostik und Lernprozessanalyse verwendet (Belusa & Eberwein, 1997). Wichtiges Kriterium ist zudem die Einbindung laufender Feedbackprozesse, die eine Rückmeldung über die Diskrepanz zwischen der aktuellen Performanz und der Zielvorgabe geben. Dies hat u.a. die Förderung des Lernerfolgs als auch der Lernmotivation zur Folge (Nicol & Milligan, 2006; Rheinberg & Krug, 2005).

3. Wechselverhältnis von Diagnose und Intervention: Diagnose und Intervention stehen in der Förderdiagnostik in einem Wechselverhältnis. Für die Ableitung von Fördermaßnahmen ist die Diagnose notwendig, um an den individuellen Leistungsstand anknüpfen zu können. Nach Einsatz der Fördermaßnahmen

wird der Erfolg der Interventionsmaßnahmen überprüft, um anhand der evtl. gesteigerten Performanz neue Fördermaßnahmen zu gestalten oder bei gleichbleibender Performanz die Fördermaßnahmen zu modifizieren. Bei mehrfachen Sequenzen von Förderung und Intervention entsteht so ein förderdiagnostischer Regelkreis (Rittmeyer, 2005, S. 18 in Anlehnung an Schönrade & Pütz, 2004).

4. Einbettung in das reale Umfeld der Teilnehmenden: Viele relevante Informationen lassen sich aus Beobachtungen in Alltagssituationen gewinnen. Indem die Informationsgewinnung im Alltag stattfindet, wird das Umfeld der Teilnehmenden berücksichtigt und in die Förderdiagnostik einbezogen, wie beispielsweise die zeitlichen Ressourcen oder die Einstellung zu Fördermaßnahmen. Förderdiagnostik ist somit nicht nur Lernprozess-, sondern auch Situationsdiagnostik (Breitenbach, 2007, S. 25; Engel, 2008, S. 34; Rittmeyer, 2005, S. 20).

5. Berücksichtigung von Stärken und Schwächen: Die Förderdiagnostik diagnostiziert neben dem Förderbedarf auch die Stärken des Lernenden, beispielsweise dessen Kompetenzen und Fähigkeiten (Eggert, 1997) und ist damit nicht nur defizit- sondern auch bzw. insbesondere stärkenorientiert.

6. Anwendung von Beobachtungsverfahren und Fehleranalysen: Instrumente für die Ermittlung des Förderbedarfs sind Beobachtungsverfahren und Fehleranalysen (Rittmeyer, 2005, S. 29). Beobachtungsverfahren haben das Ziel, große Einheiten des Verhaltens und Erlebens zu erfassen und finden meist im natürlichen Lebensumfeld der beobachteten Personen statt (Bortz & Döring, 2006, S. 322). Bei den Fehleranalysen wird die Identifikation von systematischen Fehlern fokussiert, da diese sowohl Rückschlüsse auf den Lernstand als auch Hinweise auf konkrete Ansatzpunkte zur Förderung liefern.

Im englischen Sprachraum ist für Förderdiagnostik der Begriff *Formative Assessment* gebräuchlich, wobei in der englischen Begriffsbestimmung eine weitere Dimension berücksichtigt wird: Die Zusammenarbeit zwischen Lehrenden und Lernenden sowie zwischen den Lernenden untereinander (McManu, 2008). Für eine gelingendes *Formative Assessment* soll eine partnerschaftliche Lernatmosphäre sowohl zwischen Lernenden und Lehrenden bestehen als auch zwischen den Lernenden selbst, um ihre metakognitiven Lernstrategien und somit die Reflexionsfähigkeit zu verbessern.

2.1.3 Zusammenfassender Vergleich und Akzeptanz diagnostischer Verfahren

Die Förderdiagnostik lässt sich von der selektiven Diagnostik anhand von sechs Dimensionen abgrenzen (in Anlehnung an Dluzak, Heinemann & Grotluschen, 2009).

Dimension	Selektive Diagnostik	Förderdiagnostik
Bezugsnorm	sozial	individuell
Datenhoheit	diagnostizierende Instanz	diagnostizierende Person
Konsequenz	Selektion	Anpassung der Lernangebote
Instanz	Fremdbeurteilung	Fremd-, Peer- oder Selbstbeurteilung
Perspektive	Statuserhebung	Status- oder Prozesserhebung
Zeitpunkt in Relation zum Bildungsangebot	vorlaufend und/oder nachlaufend	vorlaufend und/oder mitlaufend

Tabelle 1: Dimensionsausprägung selektive Diagnostik und Förderdiagnostik (in Anlehnung an Dluzak u. a., 2009, S. 34)

Die Förderdiagnostik findet in vielen Bereichen Anwendung (Schule, Weiterbildungseinrichtungen) und fokussiert unterschiedliche Altersgruppen (Schüler/Schülerinnen, Erwachsene). Bezüglich der Anwendung diagnostischer Verfahren waren viele Jahre die Überlegungen vorherrschend, Erwachsene im Grundbildungsbereich würden sich einer Diagnostik (auch wenn sie zur weiteren Förderung verwendet werden soll), aufgrund der Befürchtung weiterer Defiziterfahrungen, verwehren (Füssenich, 2004). Diese Annahmen werden durch aktuelle Erfahrungsberichte (Brigitte, 2004) und Studien (Nienkämper & Bonna, 2010; Schladebach, 2007) widerlegt. So wird in der Akzeptanzstudie zur Akzeptanz von Diagnostik in der Alphabetisierung (2008-2011) berichtet, dass ca. 75% der befragten Kursleiter und Kursleiterinnen angeben, Diagnosematerialien einzusetzen, deren Anwendung nicht zu lange dauert; 75% setzen Diagnosematerialien ein, bei denen das Verhältnis zwischen Aufwand und Ergebnis ausgeglichen ist (Bonna & Nienkämper, 2011, S. 46).

Dennoch ist davon auszugehen, dass gerade Erwachsene im Grundbildungsbereich aufgrund ihrer niedrigen Lese- und Schreibfähigkeiten schamhaft sind (Döbert, Hubertus & Nickel, 2000; Egloff, 1997; Füssenich, 2004; Schladebach, 2007). Betroffene berichten von einem Schamgefühl, sich einer Testsituation, beispielsweise in einer Institution wie die VHS, zu stellen. Dies hat eine niedrige Teilnehmer-/Teilnehmerinnenquote an Maßnahmen zu Alphabetisierung zur Folge: Obwohl die Anzahl funktionaler Analphabeten und Analphabetinnen in Deutschland 7,5 Millionen beträgt (Grotluschen & Riekman, 2011) nehmen nur ca. 30.000 Personen an Alphabetisierungskursen teil (Huntemann & Reichart, 2011, S. 32). Daher ist es wünschenswert, den Einstieg in die Teilnahme an Maßnahmen zur Alphabetisierung zu erleichtern (Wolf u. a., 2011, S. 127). Hilfreich kann hierbei die computerbasierte Förderdiagnostik sein, die einen flexiblen und anonymen Einsatz ermöglicht.

Zusammengefasst ist das Ziel der Diagnostik die Bewertung hinsichtlich Standards, Ziele und Kriterien. Erweiternd beinhaltet Förderdiagnostik ein Feedback, welches die Lücke zwischen dem aktuellen Fähigkeits- oder Arbeitsstand und den zu erreichenden Standard aufzeigt. Zudem impliziert die Förderdiagnostik Hinweise darauf, wie diese Lücke zwischen Leistungsstand und Zielvorgabe geschlossen werden kann. Unterscheidungsmerkmal zwischen der Selektions- und der Förderdiagnostik ist nicht das zur Diagnostik eingesetzte Instrument, sondern *wie* es eingesetzt wird.

2.2 Computerbasierte Diagnostik

Diverse Diagnostikverfahren werden computergestützt angeboten. In dem folgenden Kap. findet zunächst eine begriffliche Einordnung statt (Kap. 2.2.1). Anschließend werden die Anreicherungsmöglichkeiten durch Multimedia beschrieben (Rich E-Assessment, Kap. 2.2.2), um in den darauf folgenden Kapiteln auf die Vor- und Nachteile computerbasierter Diagnostik (Kap. 2.2.3) sowie auf die Potenziale im förderdiagnostischen Kontext (Kap. 2.2.4) eingehen zu können. Nachdem aktuelle Beispiele vorgestellt wurden (Kap. 2.2.5), werden schließlich innovative Itemformate der computerbasierten Diagnostik hinsichtlich möglicher Dimensionen erläutert (Kap. 2.2.6.1) und klassifiziert (Kap. 2.2.6.2).

2.2.1 Computerbasierte Diagnostik – Definition und Abgrenzung

Computerbasierte Diagnostik ist eine Subform des computerbasierten Assessments⁶ (engl. *Computer Based Assessment*) (vgl. z. B. Ruedel & Mandel, 2010). Jegliche computerbasierte Diagnostik ist also immer auch computerbasiertes Assessment. Um ein hinreichendes Verständnis bezüglich des Forschungsgegenstandes dieser Arbeit zu gewährleisten, wird im Folgenden zunächst der Begriff computerbasiertes Assessment erläutert, um daran anschließend den Begriff der computerbasierten Diagnostik zu erläutern.

Assessment bedeutet zunächst, mit Hilfe von statistisch erhobenen Daten Aussagen über den Wissenstand einer Person/Gruppe und über die Fähigkeiten einer Person/Gruppe treffen zu können. Die erhobenen Daten dienen der Bewertung von Kriterien hinsichtlich eines bestimmten Ziels (Crisp, 2007, S. 253). *Computerbasierte* Assessments bezeichnen spezifizierend den Einsatz von Computern zur Messung und Bewertung personaler Merkmale und beinhalten immer

⁶ Im Kontext computerbasierter Datenerhebung hat sich im deutschen Sprachraum der aus dem englischen Sprachraum stammende Begriff *Assessment* etabliert. Da in der Fachliteratur Begriffsausprägungen und –abgrenzungen anhand des Begriffs *Assessment* vorgenommen werden, wird auch in diesem Kapitel im Sinne eines präzisen und einheitlichen Verständnisses der Begriff *Assessment* verwendet.

die Nutzung von Informations- und Computertechnologie zur Erhebung der Daten und folglich der Bewertung von Kriterien. Sie umfassen den gesamten Prozess vom Aufgabendesign bis hin zur Datenspeicherung und -auswertung (Stöckberg, 2012, S. 591). Synonym verwendet werden oftmals die Begriffe elektronisches Assessment (engl. *E-Assessment*) oder technologiebasiertes Assessment (engl. *Technology Based Assessment*), wobei dies unter strenger Berücksichtigung der Bedeutungen der einzelnen Begrifflichkeiten nicht zutreffend ist. Technologiebasiertes oder elektronisches Assessment beschreibt den Einsatz jeglicher Informationstechnologie in der psychologischen und pädagogischen Diagnostik (Jurecka & Hartig, 2007, S. 37). Auch wenn in den meisten Fällen computerbasierte Assessments mit einem Desktop-Computer oder einem Laptop durchgeführt werden, muss dies nicht immer der Fall sein. Beispielsweise werden auch mobile Endgeräte wie Smartphones oder Personal Digital Assistants (PDAs) verwendet. Letztere sind z. B. von Vorteil, wenn Personen über einen längeren Zeitraum und in ihrem natürlichen Tagesablauf befragt werden sollen (Jurecka & Hartig, 2007, S. 37). Beim computerbasierten Assessment werden die Aufgaben am Computermonitor präsentiert, die Eingabe der Antworten erfolgt über die Tastatur und Maus bzw. über einen Touch Screen. Die Reaktionen werden elektronisch aufgezeichnet und meist auch elektronisch bzw. automatisiert ausgewertet sowie rückgemeldet.

Bezüglich des Technologieeinsatzes kann ein hierarchischer Bezug hergestellt werden. Die folgende Darstellung veranschaulicht die Beziehung verschiedener einsetzbarer Technologien im Assessment:

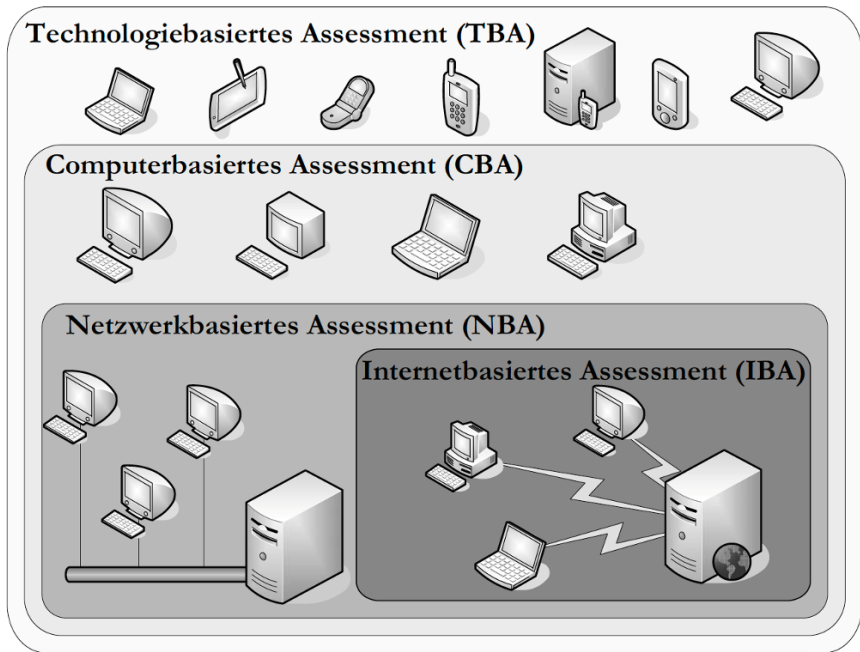


Abbildung 1: Formen des technologiebasierten Assessments (Jurecka & Hartig, 2007, S. 41)

Durchgeführt werden computerbasierte Assessments in einem online oder in einem „blended mode“. Zu computerbasierten Assessments zählen zudem internet- und/oder netzwerkbasierte Assessments. Die Unterscheidung erfolgt anhand der eingesetzten Technologie: Netzwerkbasieretes Assessment bezeichnet die Verwendung von Computernetzwerken (Jurecka & Hartig, 2007, S. 39). Diese können, müssen aber nicht unbedingt über das Internet vernetzt sein. Beim netzwerkbasierten Assessment können Tests an mehreren Computern gleichzeitig mit der Vorgabe von Testbatterien durchgeführt werden. Eine Möglichkeit dieser Form ist die Bearbeitung von Aufgaben in Gruppen. Meistens ist für ein netzwerkbasieretes Testen die Installation von Software notwendig. Im Gegensatz dazu ist die zusätzliche Installation von Software neben der Standardausstattung eines PCs oder Laptops nicht notwendig. Internetbasierte Assessments werden über das Internet und mit einem Browser durchgeführt. Der Vorteil dieser Erhebungsform ist, dass viele Teilnehmer und Teilnehmerinnen mit wenig Mehraufwand (wenn der Test einmal programmiert ist) erreicht werden können. Hier können auch Rollenspiele zur Messung der Teamfähigkeit eingesetzt werden.

Beispielsweise hat IBM Manager in ihrer Teamfähigkeit getestet und geschult, indem diese das Online-Rollenspiel World of Warcraft absolvierten (IBM, 2007). Problematisch ist allerdings die Durchführung von Leistungstests, da internetbasierte Assessments meist in unkontrollierten Umgebungen stattfinden.

Eine Unterform des computerbasierten Assessments ist die computerbasierte Diagnostik. Das Unterscheidungsmerkmal ist hierbei nicht die Technologie, sondern das Ziel der Messung, nämlich die Diagnostik. Unter der computerbasierten Diagnostik werden verschiedene Diagnoseprozesse subsumiert. Computerbasiert diagnostiziert wird beispielsweise im pädagogischen, im pädagogisch-psychologischen und im medizinischen Bereich. Für den pädagogischen Bereich werden demnach Computertechnologien für die Feststellung von Voraussetzungen und Bedingungen planmäßiger Lehr- und Lernprozesse, die Ermittlung von Lernergebnissen sowie die Analyse von Lernprozessen eingesetzt (vgl. zum Begriff Pädagogische Diagnostik Kap. 2.1.1). Ein Beispiel für den pädagogischen Bereich ist der CITO Sprachtest⁷, der in Bremen zur Sprachstandsmessung von Vorschulkindern eingesetzt wird. Dieser erhebt den Anspruch, auf der Grundlage seiner Ergebnisse, Lehrkräfte bei der Einschätzung der Schuleignung zu unterstützen⁸. Auch PISA wird seit der Erhebung 2006 zum Teil computerbasiert durchgeführt. In der Medizin werden 3D-Simulationen und interaktive Itemformate (vgl. Kap. 2.2.5) genutzt, um die diagnostische Kompetenz von angehenden Medizinern und Medizinerinnen zu fördern. Die University of Southern California führt z. B. das Projekt *Virtual Patient* durch, in welchem authentische interaktive Aufgabenformate eingesetzt werden, um klinisches Personal sowohl hinsichtlich medizinischer Diagnostik als auch Kommunikationsfähigkeiten zu trainieren⁹. Wie aus diesen Beispielen hervorgeht, bietet die computerbasierte Diagnostik Möglichkeiten (z. B. 3D-Simulationen), die über den Einsatz papierbasierter Methoden hinausgehen. In den folgenden Kapiteln werden die Anreicherungsmöglichkeiten durch Multimedia und Beispiele für computerbasierte Diagnostik vorgestellt.

⁷ www.de.cito.com (zuletzt geprüft am 03.11. 2014)

⁸ Wie aus einem Bericht der Senatorischen Behörde Bremens hervorgeht, ist es aufgrund mangelnder Computerkompetenzen (insbesondere im Umgang mit der Maus) allerdings nicht allen Kindern möglich den Test zu absolvieren (Die Senatorin für Bildung Wissenschaft und Gesundheit, 2011, S. 5). Auch wenn eine Sprachstandsfeststellung zum Zwecke der Förderung von Sprachberatern/Sprachberaterinnen und Erziehern/Erzieherinnen befürwortet wird, wird der CITO Sprachtest mehrheitlich als ungeeignet bewertet (Die Senatorin für Bildung Wissenschaft und Gesundheit, 2011, S. 8). Somit ist dieser Test und die an ihm geübte Kritik ein Beispiel dafür, dass bei der Entwicklung und dem Einsatz von Erhebungsinstrumenten die Eigenschaften potenzieller Nutzer und Nutzerinnen unbedingt zu berücksichtigen sind.

⁹ <http://ict.usc.edu/prototypes/virtual-patient/>

2.2.2 Rich E-Assessment¹⁰

„E-Assessments¹¹, die interaktive Itemformate einbinden und/oder mit auditiven und visuellen Unterstützungsfunktionen angereichert sind, werden als Rich E-Assessment bezeichnet“ (Wolf u. a., 2011, S. 129).

Der Begriff *Rich E-Assessment* hat seinen Ursprung in dem Begriff *Rich Internet Application*. *Rich Internet Application* bezeichnet Internetanwendungen, in denen Verhaltensmöglichkeiten und Funktionen von Internetanwendungen kombiniert werden und die damit aus vielfältigen Interaktionsmöglichkeiten bestehen (Colombo-Mendoza, Alor-Hernandez & Rodriguez-Gonzalez, 2012, S. 361). Erstmals verwendete Sally Jordan im Jahr 2009 eine Abwandlung des Begriffs Rich E-Assessment: Sie schrieb im Kontext von Potenzialen computerbasierter Assessments von „rich interactive e-assessment“ (Jordan, 2009). Allerdings bezog sie den Begriff ausschließlich auf interaktive Items. Merkmal von einem Rich E-Assessment ist jedoch auch die Anreicherung der Itemformate mit multimedialen Hinweisen, sogenannten *Cues* oder *Prompts* (vgl. Wolf u. a., 2011). Die visuellen oder auditiven Cues dienen der Unterstützung der Orientierung und der Navigation. Zudem kann dadurch die Aufmerksamkeit und Motivation gesteigert werden. Durch diese Unterstützungsfunktionen ist es insbesondere für Personen mit einer niedrigen ICT-Literacy möglich, an Erhebungen in unkontrollierten Umgebungen und ohne personale Unterstützung teilzunehmen (Boyle & Hutchison, 2009, S. 306).

Die Entwicklung und Etablierung von Rich E-Assessment gründete sich u. a. in den Herausforderungen, die durch die Übertragung von einer papierbasierten in eine elektronische Version entstanden. In den Anfängen des E-Assessments wurden meist papierbasierte Tests mit ihren Itemformaten 1:1 in eine elektronische Version übertragen. Dabei zeigte sich, dass die Art des Befragungsmediums einen Einfluss auf die Inhaltsvalidität haben kann: Einerseits ist zu hinterfragen, in wie weit das Itemformat direkt übertragen werden kann und den Messgegenstand evtl. verändert. Andererseits ist zu berücksichtigen, dass besondere Charakteristika – wie z. B. die ICT-Literacy – einen Einfluss auf das Messergebnis haben können (vgl. auch Kap. 2.5 zur Äquivalenzproblematik). Da E-Assessments oft auch als Self-Assessments angeboten werden, sind zudem

¹⁰ Dieses Kapitel wurde in Teilen bereits in Wolf, Koppel & Schwedes (2011) veröffentlicht.

¹¹ Da es sich um einen englischen Begriff handelt, für den es im Deutschen keine äquivalente Übersetzung gibt, wird in diesem Kapitel der englische Begriff Rich E-Assessment verwendet, wobei E-Assessment hier äquivalent mit computerbasierter Diagnostik zu verstehen ist.

motorische und physische Voraussetzungen zu berücksichtigen. Beispielsweise können Seh- und Hörfähigkeit sowie motorische Fähigkeiten einen Einfluss auf das Testergebnis haben. Diese Aspekte werden im Kontext der Äquivalenzproblematik thematisiert (vgl. Kap. 2.5). Allerdings bieten Rich E-Assessments auch gerade dahingehend Vorteile, dass Personen mit eingeschränkten physischen Fähigkeiten durch den Einsatz von Multimedia unterstützt werden können, beispielsweise durch auditive Inhalte sowie die Anpassungsmöglichkeiten der Lautstärke und Darstellungsgröße. Im Folgenden werden die Vorteile von E-Assessments im Allgemeinen und von computerbasierter Diagnostik im Besonderen erläutert.

2.2.3 Vorteile computerbasierter Diagnostik

„Technologies are well suited to supporting many of the data collection, complex analysis, and individualized feedback and scaffolding features needed for the formative use of assessment“ (Quellmalz & Pellegrino, 2009, S. 77).

Durch die technische Entwicklung unterscheiden sich inzwischen die Formate in der computerbasierten Diagnostik stark von den papierbasierten und früheren computerbasierten Formaten. Wie von Bennett erläutert und in einem späteren Artikel mit Csapó erneut erwähnt wurde, verbessert die Technologie die Diagnostik primär auf zwei Weisen: Sie verändert den Prozess an sich und sie erweitert die Messgegenstände selbst (Bennett, 2001; Csapó, Ainley, Bennett, Latour & Law, 2012). Die Verbesserung des Prozesses bezieht sich primär auf ressourcenbezogene Argumente. Die Veränderung der Messgegenstände bezieht sich auf die, mit der Weiterentwicklung der Technik auch ansteigenden, Möglichkeiten zu messender Konstrukte, so z. B. die Messung komplexer Problemlösefähigkeit mit Hilfe von Animationen und Schaubildern oder die Messung von Teamfähigkeit anhand von Rollenspielen.

Daraus ergeben sich folgende Vorteile der computerbasierten Diagnostik (in Anlehnung an Csapó u.a. (2012) sowie Pachler u. a. (2009)):

1. Flexibilität: Die Entwicklung von Tests kann mit wenig Aufwand durch einfache und/oder automatisierende Generierung von Fragen, die einfach modifiziert, verteilt und begutachtet werden können.
2. Einbindung von Multimedia: Es können dynamische Stimuli wie Audio, Video und/oder Animationen eingebunden werden, die den Einsatz von zusätzlichen Medien, wie Audioabspielgeräten, obsolet werden lassen. Zudem können moderne Technologien die Repräsentation von Ideen und die Bearbeitungen der Ideen ermöglichen und Lernende bei dem

Prozess, ihre Ideen zu präsentieren, unterstützen. Die Repräsentationsmöglichkeiten fördern wiederum die Auseinandersetzung mit den Inhalten und öffnen zudem neue Wege der Ideenfindung und –formulierung.

3. Automatisierte Auswertung und Kommunikation: Die Technologie ermöglicht die automatisierte Auswertung von Tests, die Skalierbarkeit sowie Adaptivität. Auch offene Antwortformate können automatisch ausgewertet werden. Die Ergebnisse können unmittelbar rückgemeldet werden, ohne die Notwendigkeit des Ausdrucks oder auch des Verfassens und Verschickens von Emails. Die Auswertung ist individuell oder gruppenbezogen möglich.
4. Datenspeicherung: Es können große Datenmengen gespeichert werden, die Vergleiche zwischen großen Personengruppen und/oder Längsschnittuntersuchungen erleichtern.

Die Vorteile liegen somit zum einen in der Reduzierung der Zeit- und Geldressourcen. Zum anderen in der Automatisierung und in den Gestaltungsmöglichkeiten: Es können authentische Formate erstellt werden, die u. a. eine Identifikation (deren Effekte die Motivationssteigerung und Erhaltung der Aufmerksamkeit sein können) und die Überprüfung komplexerer Fähigkeiten ermöglichen. Dies führt wiederum zur Steigerung der ökologischen bzw. externen Validität.

Trotz aller Vorteile bestehen aber auch Herausforderungen: Diese liegen in der Datensicherheit, dem Programmieraufwand, der Äquivalenz zwischen papier- und computerbasierten Befragungen und dem fraglichen Einfluss der ICT-Literacy (vgl. auch Kap. 2.5 zur Äquivalenzproblematik). Oftmals besteht eine Diskrepanz zwischen der notwendigen Erfassung personenbezogener Daten und dem Anspruch der Datensicherheit.

Speziell für die Umsetzung förderdiagnostischer Prinzipien kann die computerbasierte Diagnostik mit ihren Vorteilen gewinnbringend genutzt werden, worauf im Folgenden eingegangen wird.

2.2.4 Potenziale computerbasierter Diagnostik für die Förderdiagnostik

Unter einer computerbasierten Förderdiagnostik ist die Nutzung von Informations- und Computertechnologie für die Analyse von Kompetenzen unter Einbeziehung des bisherigen Lernverlaufs und der Identifikation von Stärken, Schwächen und der Ableitung von Fördermöglichkeiten zu verstehen (vgl. zum Begriff Förderdiagnostik Kap. 2.1.2). Somit kann die Technologie nicht genuin förderdiagnostischen Prinzipien gerecht werden – aber fast jegliche Technologie kann formativ bzw. förderdiagnostisch genutzt werden (Pachler u. a., 2009, S. 2).

Hinsichtlich der sechs Prinzipien Individualität, Prozessorientierung, Wechselverhältnis von Diagnose und Intervention, Einbettung in das reale Umfeld der Teilnehmenden, Berücksichtigung von Stärken und Schwächen sowie Anwen-

dung von Beobachtungsverfahren und Fehleranalysen (vgl. Kap. 2.1.2) lassen sich die Vorteile einer computerbasierten Förderdiagnostik im Vergleich zu einer papierbasierten Diagnostik folgendermaßen integrieren (in Anlehnung an Pachler u.a. 2009 und Kap. 2.2.3):

Bezüglich der *Individualität* bietet eine computerbasierte Förderdiagnostik die Möglichkeit der Adaptivität und eines unmittelbaren automatisierten Feedbacks. Durch die Nutzung von automatisierten Auswertungsprozessen können die Aufgaben automatisch in Abhängigkeit von bereits gelösten Aufgaben entsprechend des Kompetenzniveaus zugewiesen werden. Auch ist eine individuelle Anpassung von Inhalten (z. B. durch die Auswahl von Aufgaben in bestimmten Themenbereichen) möglich. Ebenfalls durch die automatisierte und unmittelbare Auswertung kann direkt im Anschluss der Bearbeitung einer Lernaufgabe am Computer eine Rückmeldung zur Performanz erfolgen und es können passende Übungsaufgaben zugewiesen werden. Für eine individuelle Förderung kann beispielsweise das automatische Aufrufen eines Regelkatalogs eine Unterstützung darstellen, der bei falscher Setzung von Kommata die entsprechenden Regeln aufzeigt. Auch ist ein sog. *Scaffolding*, bei dem Anleitungen, Denkanstöße und Hilfestellungen gegeben werden, um bei der Lösung von (Teil-) Aufgaben zu unterstützen, möglich (Schnotz, 2006, S. 49). Inzwischen existieren zahlreiche Studien und Einsatzszenarien über das sog. *computer embedded prompting* – das durch einen Reiz ausgelöste Erscheinen von Zusatzinformationen (Cummins, Ardeschiri & Cohen, 2008; Raes, Schellens, De Wever & Vanderhoven, 2012; Sharma & Hannafin, 2007). Die Studien belegen, dass durch den Einsatz von Scaffolding Zusatzinformationen (z. B. nach Nachschlagen von Wörtern) effektiver akquiriert werden können. Aufgrund eines geringeren Aufwands sinken zudem die Hemmschwellen, Zusatzinformationen zu nutzen (Cummins u. a., 2008, S. 17). Ferner unterstützt das computerbasierte Scaffolding nach einer Studie von Reas u.a. die Anreicherung von domänenspezifischem Wissen, das metakognitive Bewusstsein und die metakognitiven Regulationsstrategien. Allerdings ist dennoch zu erwähnen, dass der Lernerfolg mit einem durch Lehrende unterstützten Scaffolding höher ist (Raes u. a., 2012). Dies gilt insbesondere für benachteiligte Lernende, insbesondere jene mit geringem Vorwissen (Kim & Hannafin, 2011). Tendenziell geben Studien zum computerunterstützten Scaffolding Hinweise darauf, dass eine Kombination aus der Unterstützungen durch Lehrende *und* computerbasiertem Scaffolding am effektivsten für das Lernen ist (McNeill & Krajcik, 2009; Puntambekar & Kolodner, 2005; Tabak, 2004). Mit dieser Kombination können auch spezielle individuelle Bedarfe, die ggf. nicht durch ein Scaffolding abgedeckt werden können, adressiert werden.

Die *Prozessorientierung* kann computerbasiert durch automatisierte Auswertungsverfahren und die systematische Speicherung der Performanzwerte

gewährleistet werden. Das Ergebnis der Diagnostik wird gespeichert und mit Ergebnissen weiterer Durchläufe verglichen. So kann der Lernverlauf abgebildet und den Nutzern und Nutzerinnen rückgemeldet werden. Es kann berechnet werden, ob sich eine Person verbessert hat und wenn ja, in welchen Bereichen bzw. in welchem Umfang. Papierbasiert wären die drei Schritte Erhebung, Auswertung und Vergleich manuell vorzunehmen. Das hätte sowohl die Beanspruchung höherer zeitlicher Ressourcen zur Folge als auch ein höheres Fehlerrisiko. Lernstandserhebungen können somit zu mehreren Zeitpunkten durchgeführt, automatisch berechnet bzw. verglichen und den Teilnehmenden rückgemeldet werden, ohne einen hohen personalen Aufwand zu erfordern. Des Weiteren kann die Rückmeldung – insbesondere wenn sie stärkenorientiert erfolgt – zu einer Motivationssteigerung und einem größeren Lerneffekt führen (Jacobs, 2008). Die automatisierte Auswertung ermöglicht zudem eine schnelle Rückmeldung und unterstützt Lernende dabei, ggf. den nächsten Schritt unmittelbar einzuleiten, um im Problemlöseprozess voranzukommen.

Ein *Wechselverhältnis von Diagnose und Intervention* schließt an das Potenzial zur Prozessorientierung an: Computerbasiert kann eine differenzierte Rückmeldung erfolgen, die eine passgenaue Ableitung von Fördermaßnahmen ermöglicht. Eine computerbasierte Diagnostik und Förderung kann zudem adaptiv erfolgen: Die Lern- oder Diagnostikaufgaben werden je nach Performanz der vorherigen Aufgabe an das Niveau angepasst (vgl. z. B. Moosbrugger, 2008, S. 241). Des Weiteren ist eine schnelle Kommunikation mit sowohl Einzelpersonen als auch kleinen oder großen Gruppen sowie die Dokumentation des Kommunikationsprozesses möglich.

Die *Einbettung in das reale Umfeld* kann auf zwei Weisen realisiert werden: Einerseits besteht das Potenzial von neuen Medien insgesamt darin, authentische Aufgabenformate abbilden zu können (Wolf u. a., 2011). Somit kann das Umfeld der potenziellen Zielgruppe in die computerbasierte Förder- und Diagnoseinstrumente durch Simulationen (z. B. Bilder, Videos, Audio) einbezogen werden. Andererseits können mobile Endgeräte genutzt werden, die zeitlich und räumlich flexibel einsetzbar und somit an das individuelle Umfeld sowie den individuellen Tagesablauf angepasst werden können.

Die *Berücksichtigung von Stärken und Schwächen* verlangt durch die automatisierte Auswertung keinen Mehraufwand. Zeitgleich können sowohl die Kompetenzen als auch der Förderbedarf ausgewertet und dargestellt werden. Je nach Zielgruppe oder Bedarf können die Stärken und/oder die Schwächen fokussiert werden: Einerseits, indem die Rückmeldung entweder an den Bedarf angepasst ist oder die Ergebnisse von den Personen selbst ihren individuellen Bedürfnissen entsprechend abgerufen werden. Andererseits können die Diagnose-

und Lernmaterialien je nach gewünschter Fokussierung und Intensität ausgewertet werden.

Für die Anwendung von *Beobachtungsverfahren und Fehleranalysen* bietet die computerbasierte Diagnostik nur bedingt einen Vorteil. In Anlehnung an die Art und die Häufigkeit der Fehler können z. B. automatisch Rückschlüsse über systematische Fehler gezogen werden. Voraussetzung hierfür ist allerdings zum einen, dass die Diagnostik kompetenz- und lerntheoretisch fundiert ist, zum anderen, dass die möglichen Fehler bei der Diagnostik systematisiert und an das dahinter liegende theoretische Modell gekoppelt sind. Um beispielsweise das Nutzer-/Nutzerinnenverhalten hinsichtlich des Gebrauchs von Webseiten zu beobachten, können Beobachtungsverfahren mit Hilfe der Computertechnologie durch das sog. *Screenrecording* unterstützt werden, bei dem der Monitor und somit die Interaktionen zwischen Mensch und Computer aufgezeichnet werden. Die Auswertung kann teilautomatisiert erfolgen. Beim Mouse-Tracking wird beispielsweise die Bewegung der Computermaus aufgezeichnet und hinsichtlich der Zeit (Zeitspanne zwischen Interaktionen), der Verweildauer (wie lange der Mauszeiger auf einer bestimmten Stelle positioniert ist) oder auch der Häufigkeit (wie oft bestimmte Stellen bzw. Schaltflächen aktiviert werden) automatisiert ausgewertet. Ein weiteres Beispiel ist das sog. *Eye-Tracking* (vgl. Kap. 2.7.3). Tiefergehende Analysen können jedoch nur bedingt mit technologischer Unterstützung erfolgen. Beobachtungsverfahren, die komplett automatisiert ausgewertet werden, sind noch recht selten, da die Entwicklung von Algorithmen und deren Zuordnung zu Bedeutungsstrukturen aufwändig sind.

Ausblickend wurde in bisherigen Studien und Überlegungen zum computerbasierten Testen bisher wenig hinterfragt, ob die Art der Gestaltung auch einen Effekt auf das Lernen haben kann. Auch wenn beispielsweise szenariobasierte Testumgebungen eine motivationssteigernde Wirkung nachgesagt wird (s. oben) gibt es bisher wenig Belege dafür, dass computerbasierte Diagnostik nicht nur die Kompetenzen misst, sondern auch das Lernen positiv beeinflusst. Das mag einerseits daran liegen, dass die Feedbackfunktionen teilweise noch nicht ausgereift genug bzw. noch nicht vergleichbar mit menschlichen Rückmeldungen sind. Andererseits realisieren die genannten literalitätsbezogenen computerbasierten Programme zwar eine automatisierte Rückmeldung (ggf. basierend auf einem theoretischen Modell), jedoch ist diese Rückmeldung nicht immer mit einem empirisch überprüften Kompetenzmodell verknüpft. Das bedeutet, ein Nutzer/eine Nutzerin bekommt zwar eine Rückmeldung darüber, ob die Eingabe richtig oder falsch ist, jedoch keine Rückmeldungen hinsichtlich seiner/ihrer (noch zu erweiternden) Kompetenzen.

Das National Research Development Centre for adult literacy and numeracy (NRDC) in England hat einen ausführlichen Report über bestehende literalitäts-

bezogene E-Assessments (und teilweise computerbasierter Diagnoseprogramme) in England vorgelegt. Sie kommen zu dem Ergebnis, dass bestehende E-Assessments hinsichtlich ihrer Gebrauchstauglichkeit (engl. *usability*) unzureichend sind. Insbesondere gilt dies für E-Assessments, die primär von Personen mit einer niedrigen Computerkompetenz (ICT-Literacy) genutzt werden (Brooks u. a., 2005, S. 21). Folgen können mangelnde Motivation, ein dem Lernen hinderlicher *Overload* des Arbeitsgedächtnisses (vgl. Kap. 2.4) oder eine hohe Abbruchquote sein.

Zusammenfassend bietet die Computertechnologie insbesondere durch die Mobilität, die ressourcenschonende Möglichkeit automatisierter Rechenschritte und der Abbildung komplexer Inhalte Vorteile gegenüber papierbasierten Methoden. Wird eine computerbasierte Diagnostik auch als Selbstdiagnostik angeboten, besteht darüber hinaus die Möglichkeit, Hemmschwellen abzubauen: Nutzer und Nutzerinnen können in einem anonymen Umfeld ihre Kompetenzen und Fähigkeiten testen und automatisiert eine Rückmeldung erhalten. Dies kann dazu beitragen, Hemmschwellen zu senken, denn eine Testsituation in einem nicht anonymisierten Umfeld führt wiederum womöglich zu Schamgefühlen hinsichtlich der Offenbarung der eigenen Kompetenzen (vgl. z.B. Döbert u. a., 2000; Egloff, 1997; Füssenich, 2004; Schladebach, 2007). Die genannten Vorteile bezüglich der Ressourcenschonung sollen nicht suggerieren, dass eine betreuende und lehrende Person überflüssig ist. Präsenzsituationen bieten den Vorteil, flexibler auf die Teilnehmenden reagieren zu können. Zudem kann eine Präsenzsituation das Gefühl der sozialen Eingebundenheit stärken und somit zu einer Motivationssteigerung führen (Deci & Ryan, 1993; Friedrich & Mandl, 1997).¹² Zweifellos hängen aber die Vor- und Nachteile von Präsenzsituationen bzw. einem anonymen Umfeld von Testinhalt, Testsituation und der Zielgruppe ab. Es wäre wünschenswert, in Zukunft die Überlegungen über mögliche Lerneffekte durch computerbasierte Diagnostik und über eine zielgruppenspezifische Gestaltung stärker in die Forschung einzubeziehen, um die Potenziale umfassender ausschöpfen zu können.

2.2.5 Beispiele computerbasierter Diagnostik

Insbesondere in den USA ist die Entwicklung von E-Assessment-Techniken weiter vorangeschritten als in Deutschland. Im Folgenden werden Beispiele vorgestellt, in denen die Vorteile computerbasierter Diagnostik – *Flexibilität, Einbindung von Multimedia, automatisierte Auswertung und Datenspeicherung*

¹² Weiterführend sei an dieser Stelle auf Vor- und Nachteile von E-Learning und Blended Learning hingewiesen. Diese geben weitere Hinweise auf lernunterstützende und motivationssteigernde Faktoren von Präsenzsituationen im Vergleich zu mediengestützten Formaten ohne Präsenzformen (vgl. z. B. den Sammelband „Online Lernen“, herausgegeben von Klimsa, 2011).

(vgl. Kap. 2.2.3 zu Vorteile computerbasierter Diagnostik) – zum Ausdruck kommen.

Flexibilität: Ein Beispiel für die Flexibilität ist das e-asTTIE-Tool¹³. Dabei handelt es sich um ein Online Assessment Tool, welches die Messung von Schüler-/Schülerinnenleistungen sowie deren Entwicklung in Lesen, Mathematik, Schreiben und einigen wenigen Fremdsprachen ermöglicht. Lehrende können mit dem Tool die Tests selbst entwickeln, indem auf der Grundlage einer sogenannten linearen heuristischen Programmierung die passenden Itemformate gewählt und an das Curriculum und die Bedürfnisse der Lehrenden angepasst werden können. Die Ergebnisse des Online Assessments ermöglichen die Ableitung individueller Fördermaßnahmen. Die Darstellung des Lernprozesses trägt zum Verständnis der individuellen Lernentwicklung bei. Zudem können Lehrende sowohl eine fachliche Bezugsnorm wählen, welche die individuellen Leistungen mit den Anforderungen im Curriculum, als auch eine soziale Bezugsnorm, welche die Leistungen mit dem nationalen Leistungsdurchschnitt vergleicht, in Beziehung setzt. Alleinstellungsmerkmal ist neben der Möglichkeit des Vergleichs auf nationalem Niveau und der eigenständigen Entwicklung von Items auch die Bilingualität: Neben der englischen Sprache bietet das Tool auch die morische Sprache „te reo“ (polynesische Sprache des indigenen Volkes der Māori in Neuseeland) angeboten.

Einbindung von Multimedia: Durch die Einbindung von Multimedia ist inzwischen auch die computerbasierte Messung „höherer Kompetenzen“ (engl. *higher order skills*) möglich, wie die der komplexen Problemlösefähigkeit oder der Teamfähigkeit. Diese werden beispielsweise mit szenariobasierten Diagnoseprogrammen erhoben. Bei der Software *Primum Computerbased Case Simulations* (National Board of Medical Examiners (NBE), 2009) handelt es sich um ein Instrument, mit dem anhand von Szenarien die Problemlösefähigkeit von (angehenden) Medizinern und Medizinerinnen getestet wird. Jedes Szenario ist eine dynamische und interaktive Simulation eines Patienten-/Patientinnenfalls, um diagnostische Kompetenzen, Behandlungsmethoden und Begleitung zu evaluieren. Dies umfasst die Anamnese, die Diagnose und die daraus folgenden Konsequenzen (wie beispielsweise Überweisung zu anderen Fachexperten/-expertinnen). In Abhängigkeit der Handlungen verändert sich der Zustand des Patienten/der Patientin. Begonnen wird jedes Szenario mit der Darstellung des klinischen Settings, die Angabe der Zeit sowie einführende Informationen zum Patienten/zur Patientin. Die zeitliche Dimension wird ebenfalls berücksichtigt, indem Fälle mit akuten Anliegen sofort behandelt werden müssen und Fälle mit chronischen Leiden auch zeitlich verzögert behandelt können. Verwendet wer-

¹³ <http://e-asttle.tki.org.nz/> (zuletzt geprüft am 03.11. 2014)

den hauptsächlich Multiple Choice- sowie Freitext-Formate. Der Auswertungsalgorithmus wurde auf der Basis von Experten-/Expertinnenurteilen gebildet. Auch die Freitextformate werden automatisch ausgewertet – wobei die Freitextaufgaben nicht zum Schreiben langer Texte, sondern kurzer Berichte auffordern. *Primum* wird eingesetzt für die Prüfungsvorbereitung und auch Teilprüfungen für angehende Mediziner und Medizinerinnen. Für die Nutzung können ausführliche niveaudifferenzierte Manuals herangezogen werden, in denen mögliche Szenarien, Handlungsmöglichkeiten und „Spielregeln“ erläutert werden¹⁴. Bei den Szenarien muss somit nicht nur Wissen angewandt, sondern es müssen komplexe Szenarien analysiert und unterschiedlichste Informationen berücksichtigt werden, die sich in Abhängigkeit der Eingaben der Nutzer und Nutzerinnen verändern.

Im medizinischen Bereich existieren weitere Beispiele, wie die Software *Virtual Patient* der Universität Heidelberg¹⁵. Eines der wenigen Beispiele mit 3D-Simulationen ist das *Virtual Patient Project* von Mitarbeitenden der New York University, bei dem Synergien zwischen Computerspielen und medizinischer Visualisierung genutzt werden¹⁶. Zudem existiert für angehende Neurochirurgen *VCath*, eine *Virtual learning iPad App*, um berufsrelevante Fähigkeiten zu verbessern. Entwickelt wurde sie von Mitarbeitenden der Bangor University¹⁷.

Als weiteres Beispiel computerbasierter Diagnostik der letzten Jahre, in denen der Vorteile der Einbindung von Multimedia deutlich wird, ist das sog. *Recruitment*¹⁸. Dabei handelt es sich um den Einsatz spielerisch-simulativer Elemente im Personalbereich (z. B. bei der Personalauswahl). Der Begriff setzt sich zusammen aus den zwei Begriffen *Entertainment* und *Recruiting* (engl. für Personalbeschaffung). Dabei wird auf die Kombination von spielerischen Ansätzen mit Berufsorientierung, webbasiertem Personalmarketing oder auch mit eignungsdiagnostischer Auswahl via Internet gesetzt, die über die Einbindung von Medien (beispielsweise Videos und Simulationen) realisiert wird. Der Fokus liegt sowohl auf der Auswahl der Kandidaten und Kandidatinnen als auch auf der Informationsvermittlung und Unterhaltung. Für die Auswahl von Bewerbern

¹⁴ Abrufbar sind die Manuals unter <http://www.usmle.org/practice-materials/index.html> (zuletzt geprüft am 03.11. 2014)

¹⁵ <http://www.medizinische-fakultaet-hd.uni-heidelberg.de/> (zuletzt geprüft am 03.11. 2014)

¹⁶ www.tinkering.net/vp (zuletzt geprüft am 03.11. 2014)

¹⁷ <http://www.bangor.ac.uk/cs/full.php.en?nid=15617&tnid=15617> (zuletzt geprüft am 03.11. 2014)

¹⁸ Angemerkt sei, dass Recruitment sowohl als E-Assessment als auch für die computerbasierte Diagnostik eingesetzt wird. Unterscheidungsmerkmal ist dabei der theoretische und empirische Hintergrund: Liegt dem Erhebungs- und Auswertungsprozess ein theoretisch und empirisch fundiertes Modell zugrunde, handelt es sich tendenziell um Diagnostik, wohingegen ein E-Assessment diesen Anspruch nicht verfolgt (vgl. auch Kapitel 2.2 zur Definition).

und Bewerberinnen werden eignungsdiagnostische Inhalte (z. B. kognitive Leistungstests) – unter Berücksichtigung der wissenschaftlichen Gütekriterien Objektivität, Validität und Reliabilität – mit Personalmarketing-Botschaften verknüpft und in einem für die Teilnehmenden möglichst angenehmen Ambiente präsentiert (Kupka, Martens & Diercks, 2011, S. 56).

Die Mitarbeitenden der Firma Cyquest, die sich auf den Bereich Recrutainment spezialisiert hat, erachten die "Selbstausswahl der Kandidaten und Kandidatinnen gegenüber der Fremdauswahl durch die Organisationen, die Verbindung von Serious Games und Social Media als Rekrutierungskanal mithilfe von "Realistic Job Previews" sowie Facebook-Applikationen und die damit verbundene authentische Kommunikation der Arbeitgebermärkte" als Trends des Recruitments (Kupka u. a., 2011, S. 53). Die virtuellen Umgebungen können neben der Unterstützung von Unternehmen zur Auswahl von Bewerbern und Bewerberinnen auch der Selbstdiagnostik dienen. Mit einer Selbstdiagnostik kann im Anschluss an einen Durchlauf dem Teilnehmenden zurückgemeldet werden, wie er/sie im Vergleich zu anderen Teilnehmenden abgeschnitten hat. Über die virtuelle Umgebung und die Fragen wird dem Teilnehmenden zudem ein möglichst authentisches Bild des potenziellen Arbeitsplatzes vermittelt. Die möglichst authentische Darstellung des Unternehmens soll dem Teilnehmenden einen realistischen Eindruck der späteren Anforderungen vermitteln und eine möglichst umfassende Basis für die Entscheidung für oder gegen eine Bewerbung liefern. Wird den potenziellen Bewerbern und Bewerberinnen ein möglichst realistisches Bild vermittelt, können positive Effekte bezüglich der Fluktuation, der Leistungsfähigkeit oder der Abbruchquote erzeugt werden (Kupka u. a., 2011, S. 54). Ein Beispiel für den Einsatz von Recrutainment ist die Kanzlei Houthoff Burama. Die Kanzlei möchte mit dem Recrutainment-Format "Houthoff Burama - The Game" gute Absolventen und Absolventinnen für das Unternehmen interessieren und passende Bewerber und Bewerberinnen identifizieren. In dem Spiel werden potenzielle Bewerber und Bewerberinnen in die Rolle von Anwälten und Anwältinnen bei Houthoff Burama versetzt und dazu aufgefordert, an einem fiktiven aber dennoch realistischen Fall zu arbeiten. Dabei werden allerdings keine juristischen Fachkenntnisse vorausgesetzt, es werden vielmehr Problemlösekompetenzen und soziale Kompetenzen fokussiert.

Voraussetzungen für gelungene Recrutainment-Prozesse sind psychologisch fundierte Anforderungsanalysen sowie deren Operationalisierung und die zielgruppengerechte Gestaltung. Letzteres betrifft sowohl die grafische als auch die technische Umsetzung.

Automatisierte Auswertung: Beispiele für die Entwicklung im Bereich der automatisierten Auswertung sind der sog. *Language Independent Sequence*

Comparison (LISC) und der *Pearson's Versant™ Test*¹⁹. Diese analysieren und bewerten die grammatikalische Struktur von geschriebenen Texten. Die computerbasierte Diagnostik ist weniger zur Literalisierung, als vielmehr für das Lernen einer Zweitsprache gedacht. Der Versant-Test ermöglicht die Messung sowohl gesprochener als auch geschriebener Sprache hinsichtlich linguistischer Einheiten wie Textsegmente, Wörter und Silben. Die Messung der geschriebenen Texte ermöglicht zudem die Auswertung gesamter Sätze. Der LISC ist ein webbasiertes System, mit dem die Übersetzung einzelner Sätze geübt werden kann, indem Fehler erkannt und rückgemeldet werden. Das System vergleicht die eingegebenen Sätze mit möglichen korrekten Antworten hinsichtlich Rechtschreibung und Grammatik (z. B. Verwendung des richtigen Artikels im Deutschen), allerdings wird die Position der Wörter bei der Bewertung nicht berücksichtigt, sondern es wird nur gekennzeichnet, dass sich das entsprechende Wort nicht an einer korrekten Stelle befindet (Fowler, 2008). Weitere Beispiele der automatisierten Auswertung werden in den Kapiteln 2.2.6 und 2.3.6 vorgestellt und näher erläutert.

Datenspeicherung: Der Vorteil der Datenspeicherung kommt durch die prominenten Beispiele der internationalen Vergleichsstudien PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study) und PIAAC (Programme for the International Assessment of Adult Competencies) zum Ausdruck. In PISA wird das sog. TAO-System²⁰ (TAO ist das Akronym für Testing Assisté par Ordinateur, was übersetzt der Ausdruck für computerbasiertes Testen ist) verwendet. Mit Hilfe dieses Systems können große Kohorten parallel getestet und unter bestimmten Kriterien ausgewertet werden. Das System bietet die Umsetzung und das Management des gesamten Prozesses an – von der Itemerstellung bis hin zur Auswertung.

Am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) wird das TAO-System genutzt und weiterentwickelt²¹. TAO unterstützt somit die Möglichkeit der Datenspeicherung von großen Datenmengen, dessen Manage-

¹⁹ www.versanttest.com (zuletzt geprüft am 03.11. 2014)

²⁰ Entwickelt wurde das TAO-System von der Universität Luxembourg sowie dem Centre de Recherche Public Henri Tudor. Weitere Informationen sind abrufbar unter: www.taotesting.com (zuletzt geprüft am 03.11. 2014).

²¹ Am DIPF existiert seit 2007 die Forschungsgruppe „Technology Based Assessment“ (TBA), die grundlagen-, problem- und anwendungsorientierte Forschung zum Thema „Testmethoden für die Zukunft: Innovative Verfahren für die technologiebasierte Kompetenzmessung“ durchführt. Dabei werden die Bereiche computerbasiertes adaptives Testen, computerbezogene Kompetenzen (ICT-Literacy), Effekte technologiebasierter Tests, offene Textformate mit automatischer Antwortcodierung sowie technologiebasierte Diagnostik unter Berücksichtigung von Bearbeitungsschwierigkeiten fokussiert. Eines der Ziele ist, einen internationalen Standard für computerbasiertes Testen zu etablieren.

ment und dessen Auswertung. Es ist zu erwarten, dass durch die Komprimierung von Daten immer weniger Speicherplatz für die Datenspeicherung benötigt wird. Damit steigen auch die Möglichkeiten, Testergebnisse großer Kohorten über einen längeren Zeitraum speichern und auswerten zu können.

Die angeführten Beispiele zeigen lediglich einen kleinen Ausschnitt und geben einen Eindruck davon, welche Möglichkeiten in der computerbasierten Diagnostik (und auch Förderdiagnostik) bestehen. Insbesondere die Optionen zur Simulation, zur Darstellung komplexer Probleme oder dynamischen Wechselwirkungen sowie zur automatisierten Auswertung bringen die Vorteile der computerbasierten Diagnostik zum Ausdruck und stellen entscheidende Unterscheidungsmerkmale und den Mehrwert im Vergleich zur papierbasierten Diagnostik dar. In Bezug auf förderdiagnostische Potenziale dieser Beispiele sei angemerkt (wie bereits eingangs kurz erwähnt), dass es sich bei diesen Beispielen nicht genuin um Förderdiagnoseinstrumente handelt. Weder liefern die Online Tools Ergebnisse zu möglichen Lernstrategien, Lernwiderständen oder systematischen Fehlern (vgl. Kap. 2.1.2), noch werden passende Fördermaßnahmen von dem Systemen vorgeschlagen. Somit können die genannten Beispiele zur Umsetzung von Förderdiagnostik beitragen, nicht jedoch insgesamt als Förderdiagnoseinstrumente bezeichnet werden.

Zweifellos wird der Bedarf an computerbasierter Diagnostik – insbesondere durch groß angelegte internationale Vergleichsstudien wie beispielsweise PISA – und an der Forschung zu diesem Thema bestehen bleiben bzw. weiterhin ansteigen. Mit der (Weiter-) Entwicklung von Technologien und Software steigen die Möglichkeiten zur Entwicklung von Testumgebungen und -szenarien. In den bisherigen Kapiteln nicht explizit betrachtet wurde das Items, ein unabdingbarer Bestandteil eines Tests. Mit der technischen Entwicklung steigen somit auch die Potenziale für die Umsetzung von Itemformaten. Itemformate, in denen technische Neuentwicklungen umgesetzt oder auch auf eine neue Weise miteinander verknüpft werden, werden auch als innovative Itemformate bezeichnet (Parshall, Davey & Pashley, 2000) (vgl. folgendes Kap.). Die angeführten Beispiele haben bereits darauf hingedeutet, dass die Umsetzung computerbasierter Diagnostik je nach Messgegenstand und -absicht variiert. In Zusammenhang mit dem Messgegenstand und der Messabsicht stehen die Auswahl und Gestaltung der Itemformate, mit denen die Daten für die Diagnostik erhoben werden. In den vorherigen Kapiteln wurde auf die computerbasierte Diagnostik im Allgemeinen eingegangen, in den folgenden Kapiteln sollen nun im Besonderen die Charakteristiken innovativer Itemformate (die zum Großteil auch in den eben angeführten Beispielen Anwendung finden) vorgestellt und näher erläutert werden.

2.2.6 Innovative Itemformate

Ein Itemformat ist eine Einheit, welche die Wahl einer oder mehrerer Antwortmöglichkeiten erfordert. Die Antwort wird nach festgelegten Regeln bewertet (Haladyna, 2004, S. 41). Ein Item besteht aus den Komponenten Frage/Aufforderung, Bedingungen für die Angabe der Antwort sowie eine Auswertungsprozedur (Auswertungsalgorithmus). Es gibt verschiedene Formen von Items, wie z. B. offene Itemformate (z. B. Freitexte) und geschlossene Itemformate (z. B. Single Choice).

Mit der Weiterentwicklung von Technologien eröffnen sich neue Räume für die Darstellung und Nutzung von Itemformaten. Im Bereich der computerbasierten Diagnostik werden Items, die technologische Möglichkeiten des Testens (nicht nur bezogen auf Visualisierungsmöglichkeiten, sondern insbesondere auch neue Möglichkeiten der Messung sowie Auswertung) nutzen und die nicht mit konventionellen Testverfahren möglich wären, als innovative Itemformate bezeichnet (Parshall u. a., 2000, S. 129). Dabei geht es nicht nur darum, existierende Verfahren mit neuen Technologien anzureichern, sondern die Technologie dazu zu nutzen, um entweder den *Prozess zu verändern* oder aber auch die möglichen *Messgegenstände zu erweitern* (Bennett, 2001, S. 1). Einerseits steigert die Entwicklung von Technologien die Möglichkeiten einer automatisierten Auswertung. Andererseits ermöglichen neue Technologien auch die Messung von Konstrukten, die vorher nicht oder nur sehr aufwendig gemessen werden konnten (Csapó u. a., 2012, S. 153). Innovative Itemformate können sich daher durch die Erweiterung der *Messgegenstände* und/oder der Automatisierung des *Auswertungsprozesses* auszeichnen.

Hinsichtlich der erstmalig möglichen Messung von Konstrukten nennen Csapó u. a. (Csapó u. a., 2012, S. 153) die Beispiele Problemlösen in technologiereichen Umgebungen (Bennett, Persky, Weiss & Jenkins, 2010), analytische und dynamische Aspekte von Problemlöseprozessen (Wirth & Klieme, 2003), Teamwork mit Hilfe von sog. *Situational Judgement Tests* (SJTs)²² (Kyllonen, 2009) sowie dynamische Entwicklungen anhand der Einbeziehung des Prozesses im Kontext des computerbasierten Assessments (Ainley, 2006; Hadwin, Winne & Nesbit, 2005).

Bezüglich des automatisierten *Auswertungsprozesses* erlaubt die computerbasierte Messung inzwischen auch die Auswertung offener Antwortformate. Auf die Messung von beispielsweise Literalität bezogen, können neben Wörtern oder einzelnen Sätzen inzwischen auch gesamte Aufsätze ausgewertet werden (s. Beispiele computerbasierter Diagnostik in Kap. 2.2.5 und Dimensionen innova-

²² Wobei anzumerken ist, dass auch internetbasierte Computerspiele wie beispielsweise World of Warcraft genutzt werden, um die Teamfähigkeit zu prüfen (IBM, 2007).

tiver Itemformate in Kap. 2.2.6.1). Lückentexte oder offene Formate (wie beispielsweise ein Aufsatz) sind momentan den innovativen Formaten zuzuordnen²³, da die Auswertung solcher Formate komplexe Berechnungen erfordert. Der bereits in Kap. 2.2.5 beschriebene *Pearson's Versant™ Tests*²⁴ misst gesprochene Sprache und das Leseverstehen. Der Auswertung liegt ein linguistisches Modell zugrunde, wobei im Detail die Kompetenzbereiche Aussprache, Struktur, Vokabular, Sprachfluss, Verständnis und Interaktion geprüft werden. Die automatisierte Auswertung der Diagnosesoftware wurde auf der Grundlage vieler Daten in Form von gesprochenen Texten (inklusive Nicht-Muttersprachlern, um eine möglichst große Bandbreite an Akzenten und Dialekten berücksichtigen zu können) und deren Auswertung durch Experten und Expertinnen realisiert (Ripley, Tafler, Ridgway, Harding & Hakan, 2009). Ein ähnliches Vorgehen wurde auch bei dem sogenannten *Pearson Reader*²⁵ gewählt, der die automatisierte Auswertung von Aufsätzen ermöglicht. Im anschließenden Kapitel wird das Beispiel der Literalitätsmessung mit innovativen Itemformaten hinsichtlich des Auswertungsalgorithmus⁶ erneut aufgegriffen.

Bereits diese Erläuterung innovativer Itemformate deutet darauf hin, dass diverse Itemformate mit unterschiedlichen Ausprägungen (z. B. offene und geschlossene Formate) und Gestaltungsmöglichkeiten (z. B. durch die Einbindung von Multimedia) existieren. Entscheidend für die Umsetzung und die Auswahl von Itemformaten ist, was gemessen werden soll: Geht es um fachspezifisches Wissen, Teamfähigkeit oder komplexe Problemlösefähigkeit? Abhängig von dem Messgegenstand sind unterschiedliche Itemformate mehr oder weniger geeignet. Zudem bestimmt das Itemformat die Auswertungsmöglichkeiten. Diese können einfach (z. B. Multiple Choice-Formaten) oder komplex (z. B. offene oder dynamische Antwortformate) sein, wobei letztere nur mit hohem Programmieraufwand automatisiert ausgewertet werden können. Studien belegen, dass Itemformate mit hoher Interaktivität, mit der Möglichkeit Konstruktionsleistungen zu erbringen und den Lernprozess selbst zu kontrollieren sowie zu evaluieren, einen motivationssteigernden Effekt haben (Ally, 2004; S. K. Reed, 2006; Svinicki, 1999).

Die Auswahl passender Itemformate für die Messabsicht kann durch Taxonomien (Klassifikationsschemata) unterstützt werden. Hierfür werden im folgenden Kapitel Dimensionen innovativer Itemformate erläutert und anschließend taxiert.

²³ Die Bewertung „innovativ“ ist natürlich immer im zeitlichen Kontext in Abhängigkeit des aktuellen Entwicklungsstands zu betrachten.

²⁴ www.versanttest.com (zuletzt geprüft am 03.11. 2014)

²⁵ www.pearson.com.au/educator/secondary/digital-learning/pearson-reader-20/ (zuletzt geprüft am 03.11. 2014)

2.2.6.1 Dimensionen innovativer Itemformate

In bestehenden Taxonomien werden Itemformate anhand unterschiedlicher Dimensionen klassifiziert. Die Taxonomien unterstützen damit die dem Messgegenstand angemessene Auswahl passender Itemformate und können vordergründig zwei Organisationsschemata zugeordnet werden: *Messung* und *Computertechnologie* (vgl. Wolf u. a., 2011). Taxonomien, in denen die Möglichkeiten der (diagnostischen) Messung – z. B. die Offenheit des Antwortformates sowie die dem Format entsprechende Auswertungsweise – fokussiert werden, sind dem Organisationsschema *Messung* zuzuordnen (vgl. z. B. Taxonomien von Bennett, 1993; Scalise & Gifford, 2006; Snow, 1993). Taxonomien, die vorwiegend Formate hinsichtlich der eingesetzten Computertechnologie unterscheiden, zählen zum Organisationsschema der *Computertechnologie* (vgl. z. B. Taxonomien von Koch, 1993; Parshall & Harmes, 2007).

In bisher bestehenden Taxonomien wird meist nur eines der beiden Organisationsschemata berücksichtigt (eine Ausnahme stellt die Taxonomie von Parshall, Cavey und Pashley (2000) dar). Darüber hinaus werden die Möglichkeiten zur Einbindung von Medien (zum Beispiel Animationen, Filme, Bilder etc.) nur bedingt systematisch integriert.

In einem Literatur-Review von Wolf u. a. wurden aussagekräftige Taxonomien zu Itemformaten (Boyle & Hutchison, 2009; Parshall u. a., 2000; Scalise & Gifford, 2008) gesichtet (Wolf u. a., 2011)²⁶. Auf der Literaturbasis und unter Berücksichtigung aktueller Entwicklungen in der computerbasierten Diagnostik wurde eine Taxonomie entwickelt, in der sowohl das Organisationsschema Messung als auch Computertechnologie integriert sind. Die Taxonomie ist stark an die von Parshall u. a. (2000) angelehnt, welche aus den Dimensionen Itemformat, Antwortaktivität, Medienanreicherung, Interaktivität und Auswertungsalgorithmus besteht. Die Taxonomie von Wolf u. a. (2011) besteht ebenfalls aus fünf Dimensionen, die allerdings nicht deckungsgleich mit der Taxonomie von Parshall u. a. (2000) sind:

²⁶ Das in diesem Kapitel behandelte Thema wurde bereits in leicht veränderter Form in Wolf, Koppel & Schwedes (2011) veröffentlicht. Die hier vorgestellte Taxonomie weicht in Teilen von den bereits veröffentlichten Inhalten ab. Begründet ist dies darin, dass die Items und die Online-Testumgebung weiterentwickelt und die Bewertung der Itemformate an den Entwicklungsstand angepasst wurden.

1. Offenheit des Antwortformats
2. Auswertungsalgorithmus
3. Authentizität
4. Medienanreicherung
5. Interaktivität

In der Taxonomie von Wolf u. a. (2011) ist im Vergleich zu der von Parshall u. a. (2000) die Dimension *Antwortaktivität* nicht explizit berücksichtigt und sie ist um die Dimension der *Authentizität* erweitert. In den Dimensionen sind beide Organisationsschemata berücksichtigt: *Offenheit des Antwortformats*, *Auswertungsalgorithmus* sowie *Authentizität* sind dem Schema Messung zuzuordnen. *Medienanreicherung* und *Interaktivität* kategorisieren hingegen Gegenstände der Computertechnologie. Die Innovativität stellt eine zusätzliche Dimension dar, die zu den eben genannten Dimensionen „quer“ liegt, d.h. jede der fünf Dimensionen weist neben ihren dimensionsspezifischen Ausprägungen zudem eine Ausprägung in der Dimension *Innovativität* auf. Wie im Folgenden deutlich werden wird, sind die fünf Dimensionen in ihrer Umsetzung nicht gänzlich unabhängig voneinander zu betrachten (so beeinflusst beispielsweise die Offenheit des Antwortformats auch den Auswertungsalgorithmus). Damit hat mitunter ebenfalls die Ausprägung der Innovativität in einer Dimension eines Items einen Einfluss auf den innovativen Charakter in einer weiteren Dimension des Items (z. B. kann eine starke Medienanreicherung auch eine höhere Interaktivität zur Folge haben).

Die Dimensionen werden, unter Zuordnung zu den Organisationsschemata Messung und Computertechnologie, im Hinblick auf ihre Ausprägungsmöglichkeiten und auf den Einsatz in computerbasierter Diagnostik erläutert, um im anschließenden Kapitel die Dimension der Innovativität hinzuzuziehen.

Organisationsschema Messung

1. Offenheit des Antwortformats: Die Dimension Offenheit des Antwortformats beschreibt, wie offen die Handlungsspielräume für die Teilnehmenden bei der Eingabe der Antwort sind. Die Dimension ist als ein Kontinuum zu betrachten, welches im Folgenden nach Scalise und Gifford differenziert wird (Scalise & Gifford, 2006). Als „geschlossenste“ Antwortformate gelten *Multiple Choice*-Formate, bei denen nur eine Antwort (z. B. richtig/falsch oder ja/nein) ausgewählt werden kann (Scalise & Gifford, 2006, S. 11). Allerdings ist anzumerken, dass die Kategorie MC-Formate häufig synonym auch für *Single-Choice*-Formate verwendet wird. Bei Single Choice-Formaten kann lediglich eine Antwort, bei Multiple Choice-Formaten können hingegen auch mehrere Antworten

ausgewählt werden (Haladyna, 2004; Scalise & Gifford, 2006) – worauf auch die Übersetzung von *multiple choice* mit „Mehrfachauswahl“ hinweist.

Die nächste Kategorie auf dem Kontinuum Richtung Offenheit sind halboffene Antwortformate. Zu ihnen zählen *Selection-/Identification*-Formate. Diese Kategorie umfasst *Complex Multiple Choice* bzw. *Multiple Answers*. Hier stehen mehrere Antworten zur Auswahl. Die Distraktoren (falsche Antwortmöglichkeiten) bieten dabei das Potenzial, die Ratewahrscheinlichkeit zu minimieren. Diese Formate, wie sie von Scalise und Gifford definiert werden, werden oftmals als Multiple Choice-Formate beschrieben. Abweichend von der vorgeschlagenen Differenzierung von Scalise und Gifford erscheinen die Kategorien Single Choice (an der Stelle von dem Multiple Choice-Formaten nach dem Verständnis von Scalise und Gifford) und Multiple Choice-Format (an der Stelle von *Selection/Identification*) brauchbarer und verständlicher. Denn der Terminus *Multiple* beim Multiple Choice-Format deutet darauf hin, dass mehrere Möglichkeiten zur Auswahl stehen und sinnvollerweise in Abgrenzung zum Single Choice-Format auch mehrere Antworten gewählt werden können.

Das Format *Reordering/Rearrangement* ist ebenfalls halb offen, es bestehen allerdings mehrere Antwortmöglichkeiten bzw. Kombinationsmöglichkeiten. Dabei werden die Teilnehmer und Teilnehmerinnen aufgefordert, Begriffe zu kategorisieren oder etwas an- bzw. zuzuordnen.

Darauf folgt die Kategorie *Substitution/Correction*, in denen ein Wort korrigiert oder aus bestehenden Elementen eine Figur konstruiert werden soll. Die Korrekturmöglichkeiten sind meist festgelegt, z. B. auf die Buchstaben des Alphabets.

Ein weiteres halboffenes Format ist *Completion*, welches dazu auffordert, kurze Antworten einzutragen, Sätze zu komplettieren, Lückentexte auszufüllen oder Matrizen zu ergänzen. Die Antwortmöglichkeiten (z. B. die Möglichkeiten von Schreibweisen) sind dabei vielzählig (je nach Anzahl der Wörter).

Das Format *Construction* beinhaltet Aufgaben, in denen Lösungen konstruiert werden. Die Konstruktion findet allerdings über die Anwendung einer Methode statt, beispielsweise der Erstellung eines Aufsatzes oder einer Concept Map.

Am anderen Ende des Kontinuums befindet sich das offene Format *Presentation/Portfolio*. Bei diesem Format können Präsentationen, Demonstrationen eines Experiments, Diskussionen oder auch Interviews und die Sammlung/Reflexion von Artefakten für ein (Projekt)Portfolio zur Bewertung herangezogen werden.

Geschlossene Antwortformate sind die am meisten verwendeten Formate in der computerbasierten Diagnostik (Stöckberg, 2012, S. 599). Mit geschlossenen Antwortformaten werden meist Wissen und kognitive Fähigkeiten gemessen; die

Auswertung ist objektiv. Die Schlussfolgerungen beziehen sich meist auf eine Wissensdomäne, Fähigkeiten oder beides (Haladyna, 2004, S. 47). Im Kontext der Förderdiagnostik können mit geschlossenen Formaten einfache Basiskonstrukte getestet werden, um „systematische Misskonzeptionen differenziert auf der Teilkonstruktebene zu diagnostizieren“ (Wolf u. a., 2011, S. 131). Offene Formate können für die Messung komplexer Fähigkeiten oder Kompetenzen – z. B. Schreibkompetenzen – genutzt werden.

2. Auswertungsalgorithmus: Der Auswertungsalgorithmus ist anhand zweier Kriterien zu beschreiben: Automatisierungsgrad und Auswertungsschema. Er beschreibt, in wie weit die Auswertung – bei computerbasierten Tests – automatisiert, teilautomatisiert oder nicht automatisiert erfolgt und welches Auswertungsschema angewendet wird.

Hinsichtlich der Automatisierung ist prinzipiell davon auszugehen, dass eine Antwort umso leichter automatisch ausgewertet werden kann, je geschlossener das Aufgabenformat ist. Bezüglich komplexer und offener Aufgabenformate ist die automatische Auswertung nur bedingt bzw. mit hohem Aufwand möglich. Es gibt aber mittlerweile Programme, die vollautomatisierte Auswertungen sowohl von Kurzantworten, (z. B. dem c-rater (Sukkarieh & Stoyanchev, 2009)) als auch von Aufsätzen ermöglichen (z. B. mit dem e-rater (Burstein u. a., 1998)). Allerdings bedarf eine automatisierte Auswertung von Aufsätzen einer umfangreichen Vorarbeit. Für das Training einer solchen Software und die Entwicklung automatisierter Abläufe zur Auswertung sind 200 bis 250 Aufsätze erforderlich, die von Personen „händisch“ hinsichtlich formaler Anforderungen (Grammatik, Vokabular, Sprachauswahl, Rechtschreibung, Struktur und Kohärenz) ausgewertet und in den Auswertungsalgorithmus einbezogen werden (Streeter, Bernstein, Foltz & Donald, 2011, S. 16). Anhand der festgelegten Auswertungskriterien werden die Aufsätze bewertet. Aufgrund der festgelegten Auswertungskategorien können daher auch nur die bereits verwendete Kriterien Anwendung finden. Die Programme zur automatisierten Auswertung offener Antwortformate werden somit anhand von menschlichen Bewertungen „trainiert“. Eine inhaltliche Analyse können sie nicht leisten.

Vorteile einer vollautomatisierten Auswertung sind die Möglichkeiten, adaptiv zu testen, unmittelbare Rückmeldungen einzubinden (vgl. auch Bemerkung zu schnellere Rückmeldung bezüglich der Dimension Medienanreicherung) und individuelle Fördermaßnahmen abzuleiten. Für das adaptive Testen ist allerdings Voraussetzung, dass die Items in ihrer Schwierigkeit hierarchisch (im Idealfall auf empirischer Basis) eingestuft werden, so dass je nach Kompetenzstand die entsprechenden Items zugewiesen werden können. Die beiden zuletzt genannten Möglichkeiten sind insbesondere für die Erfüllung förderdiagnostischer Prinzipien notwendig (vgl. Kap. 2.1.2).

Neben dem Grad der Automatisierung ist das zweite Kriterium vom Auswertungsalgorithmus das Auswertungsschema, welches dichotom, polytom oder komplex sein kann (Parshall & Harnes, 2007, S. 13). Dichotom bedeutet, es gibt nur richtige oder falsche Antworten. Bei polytomen Items ist die Vergabe von Teilpunkten möglich. Dieses Auswertungsschema wird auch als *partial credit* bezeichnet. Dadurch wird zudem eine unterschiedliche Gewichtung von Messkriterien ermöglicht. Komplexe Auswertungsschemata werden herangezogen, wenn eine große Variation bezüglich der möglichen Antworten besteht. Hierzu zählen beispielsweise auch offene Antwortformate wie Aufsätze. Oftmals sind solche Formate kontextualisiert und mit Unterstützung von Stimuli umgesetzt. Die automatisierte Bewertung solcher Formate ist aufwändig umzusetzen, da durch die große Bandbreite möglicher Antworten zumeist unterschiedliche Auswertungsregeln angewendet werden müssen (beispielsweise schlagen sich bei einer Rechtschreibprüfung die unterschiedlichen Regeln auch in dem Auswertungsalgorithmus nieder). Vorteilhaft ist ein komplexer Auswertungsalgorithmus allerdings dahingehend, dass durch die Berücksichtigung verschiedener Regeln (z. B. nicht nur „richtig“ und „falsch“, sondern auch die Abfolge von Entscheidungen und Gewichtungen dieser Entscheidungen) mehr Informationen über die getestete Person eingeholt werden können.

Hinsichtlich eines förderdiagnostischen Anspruchs sind insbesondere dichotome Auswertungsschemata sinnvoll, da diese am einfachsten eine dimensionsreine Erhebung und Bewertung von Kompetenzen ermöglicht.

3. Authentizität: Eine Ausprägung von Authentizität ist nach Bennett (1993) der Grad der Kontextualisierung (Bennett, 1993, S. 2). Gulikers u.a. differenzieren den Ansatz von Bennett weiter aus und beschreiben Authentizität als ein relatives Konzept:

„[...] authenticity of something can only be defined by its resemblance to something else and it is the specification of this something else that is crucial for further discussion about and examination of the concept of authenticity“ (Gulikers, Bastiaens, Kirschner & Kester, 2008, S. 22).

Ihnen zu Folge ist Authentizität nicht objektiv messbar, sondern wird subjektiv wahrgenommen. Der Grad der Authentizität hängt von den folgenden fünf Charakteristika ab: a) der Art der Aufgabe, b) dem physikalischen Kontext, in dem die Befragung eingebettet ist, c) dem sozialen Kontext des Assessments, d) dem Ergebnis bzw. der Form, in welcher das Ergebnis präsentiert wird sowie e) den Bewertungskriterien. Um eine möglichst hohe ökologische Validität zu gewährleisten, sollten in der Gestaltung authentischer Aufgabenformate das Wissen

sowie die Fähigkeiten realer Aufgaben, welche repräsentativ für das jeweilige „Testfeld“ (im beruflichen Kontext z. B. das Arbeitsfeld) sind und kultur- sowie communityspezifische Praktiken abbilden, berücksichtigt werden (J. S. Brown, Collins & Duguid, 1989). Zudem sollten die Aufgaben die Teilnehmenden dazu auffordern, ihr Wissen, ihre Fähigkeiten und Haltungen so einzubringen, wie sie in einem professionellen Kontext/Arbeitsumfeld von entsprechenden Berufstätigen genutzt und eingesetzt werden würden (J. S. Brown u. a., 1989; Darling-Hammond & Snyder, 2000; Gielen, Dochy & Dierick, 2003; J. Herrington & Oliver, 2000). So sollten auch im Sinne einer möglichst hohen prognostischen Validität (verstanden als Voraussage für die zukünftige erfolgreiche, z. B. berufliche Tätigkeit) die Aufgaben weitestgehend der künftigen Tätigkeit entsprechen. Die Bewertungskriterien sind entsprechend der geforderten Ergebnisse in der Praxis zu gestalten (Darling-Hammond & Snyder, 2000). Begründet ist diese Ausdifferenzierung in fünf Charakteristika darin, dass der Grad der Authentizität durch Beeinflussung der einzelnen Charakteristika und den Grad der Kontextualisierung verändert werden kann. Die durch die Authentizität entstandene Komplexität hat wiederum einen Einfluss auf das Frageformat und begründet somit die Zuordnung der Authentizität zum Organisationsschema der Messung. Diese Erläuterung macht die Grenze zur Interaktivität (siehe weiter unten) deutlich: Z. B. kann ein Formular hochgradig kontextualisiert und somit authentisch sein; es fordert aber keine Interaktivität über das Ausfüllen der Formularelemente hinaus. Hingegen kann eine hoch interaktive Simulation vollkommen abstrakt und somit nicht authentisch sein.

Allerdings ist gerade in der erwachsenengerechten Diagnostik aus Motivationsgründen auf die Verwendung eines arbeits- und lebensweltbezogenen Kontext zu achten. Rich E-Assessments (vgl. Kap. 2.2.2) bieten aufgrund der Möglichkeit der Einbindung verschiedener Medien ein großes Potenzial.

Organisationsschema Computertechnologie

4. Medienanreicherung: Die Medienanreicherung beschreibt die Anzahl sowie die Daten- und Informationsdichte der Medienformate, die bei der Gestaltung einer Aufgabe genutzt werden. Sie ist dem Organisationsschema Computertechnologie zuzuordnen. Im Vergleich zur Media-Richness-Theorie²⁷ von Daft und Lengel (Daft & Lengel, 1986) geht es hierbei allerdings nicht um eine Passung zwischen Kommunikationsanlass und Medienwahl sondern vorwiegend um deren Varianz. Nach Dennis und Valacich wird der Reichtum eines Mediums wie folgt beschrieben:

²⁷ In der Media-Richness-Theorie wird der Frage nachgegangen, wie eine effektive Kommunikation und Kooperation durch eine angemessene Medienwahl unterstützt werden kann.

„Richer media were those with a greater language variety (the ability to convey natural language rather than just numeric information), a greater multiplicity of cues (the number of ways in which information could be communicated such as the tone of voice), a greater personalization (ability to personalize the message), and more rapid feedback” (Dennis & Valacich, 1999).

Reiche Medien bieten demnach die Möglichkeiten einer Varianz der *Inhaltsdarstellung* (z. B. Bild- und Textsprache), der *Medieneinbindung* (Bild, Ton etc.), der stärkeren *Personalisierung* sowie der *schnelleren Rückmeldungen* (vgl. hierzu Dimension Auswertungsalgorithmus²⁸).

Inhaltsdarstellung: Inhalte können mit Hilfe der Computertechnologie auf verschiedene Weisen dargestellt bzw. codiert (Weidemann, 2002, S. 47) werden. Unterschieden wird dabei zwischen Text, Bild und Zahlen. Wird der Inhalt nur auf eine Weise dargestellt, handelt es sich um eine monocodale Darstellung – z. B. wenn der Inhalt nur über die Textebene präsentiert wird. Werden hingegen mehrere Darstellungsoptionen genutzt, liegt eine multicodale Inhaltsdarstellung vor. Über die Computertechnologie können Inhalte unterschiedlich und/oder parallel codiert werden. Studien zeigen, dass die Verwendung unterschiedlicher Codierungen je nach Kombinationen mit der Darstellungsmodalität einen positiven oder aber auch negativen Effekt auf den Lernprozess haben können. So können Inhalte beispielsweise besser aufgenommen werden, wenn eine Kombination zwischen Grafik und Audio vorliegt, im Vergleich zu einer gleichzeitigen Präsentation von Grafik und Text (vgl. Low & Sweller, 2005; Weidemann, 2002).

Medieneinbindung: Computer ermöglichen die Einbindung und Kombination unterschiedlicher Medien, wie Grafiken, Audio, Video und Animationen (Parshall u. a., 2000). Grafiken sind dabei die gebräuchlichste Form der nicht textbasierten Darstellung. Auch in papierbasierten Tests könne Grafiken eingebunden werden, doch ist es mit Hilfe von Computern und der Verbindung mit Interaktivität möglich, Grafiken zu bearbeiten, sie rotieren zu lassen und sie zu vergrößern oder zu verkleinern.

Audio wird insbesondere für die Messung musikalischer und sprachlicher Fähigkeiten (Hörverständnis) verwendet. Es können zwar auch Audioabspielgeräte in Verbindung mit einer papierbasierten Diagnostik verwendet werden, doch erfordert dies die zusätzliche separate Nutzung von Abspielgeräten, die aber nicht direkt mit dem Testinhalt verknüpft sind. Auf computerbasierter Basis

²⁸ Hinsichtlich der Medienanreicherung wird die „schnellere Rückmeldung“ nicht weiter berücksichtigt, das sich die unmittelbare Rückmeldung aus dem Grad der Automatisierung ergibt.

können Darstellungen direkt mit Audio verknüpft werden, indem z. B. auf dem Bildschirm ein Symbol oder Bereich mit der Computermaus angeklickt und die dazugehörige Audiodatei abgespielt wird. Auch ermöglicht die Einbindung von Audio die Nutzung von Computertechnologie für Personen mit geringen Lese- oder Sehfähigkeiten. Zudem kann die Einbindung von Audio das Arbeitsgedächtnis entlasten: Das Modalitätsprinzip besagt, dass – im Vergleich zu einer Inhaltsdarstellung in Textform – die auditive Darstellung in Kombination mit einer graphischen Darstellung das Arbeitsgedächtnis entlastet und somit Lernen unterstützt (Low & Sweller, 2005).

Videos ermöglichen die Darstellung von Inhalten, die papierbasiert nur schwer oder gar nicht abbildbar sind, beispielsweise komplexe Sachverhalte, die authentische Darstellung von Prozessen (z. B. Produktionsprozessen) oder von Personen (z. B. im medizinischen Bereich die Darstellung eines Patienten-/Patientinnenfalls). Wie auch die Nutzung von Audio kann Video ebenfalls im Kontext papierbasierter Testung genutzt werden. Doch liegt der Mehrwert der integrierten Nutzung von Computer und Video insbesondere darin, das Video individuell abspielen, anhalten und/oder wiederholen zu können. Gleichzeitig ist aber zu beachten, dass die Decodierung eines Videos, in dem wenig komplexe Inhalte dargestellt werden, eine erhöhte Verarbeitungsleistung im Vergleich zu einer weniger medienreichen Darstellung (z. B. einem Bild) erfordert.

Animationen haben im Vergleich zum Papiermedium den Vorteil, dass sie nicht statisch sind. Sie stellen insbesondere einen Mehrwert durch die Möglichkeit dar, komplexe dynamische Prozesse darzustellen. Animationen können gegenüber Videos vorteilhaft sein, da sie die schematische Abbildung von Prozessen und/oder auch nur einen Teilausschnitt dieser ermöglichen. So ist auch die Fokussierung von bestimmten Aspekten möglich. Des Weiteren ist für die Darstellung von Animationen weniger Rechenkapazität notwendig als für die Darstellung von Videos (Parshall u. a., 2000, S. 139).

Der Stimulus eines Mediums kann sich zudem auf weitere menschliche Sinne beziehen. Mit *tangible Interfaces* ist es möglich, einen haptischen Reiz auszulösen – so werden z. B. in der Medizin Puppen verwendet, in denen der Puls simuliert wird. Auch der Geruchssinn kann mit Hilfe von Geruchssynthesizern stimuliert werden.

Personalisierung: Inhalte, Medien und Rückmeldungen können an die jeweiligen Bedürfnisse angepasst werden. Die inhaltliche Anpassung erfolgt z. B. in adaptiven Tests, indem in Abhängigkeit der Beantwortung eines Items das darauf folgende Item präsentiert wird (vgl. auch Kap. 2.2.3 zu Vorteilen computerbasierter Diagnostik). Die eingesetzten Medien können an die individuellen Bedürfnisse angepasst werden: Bei Audios und Videos durch die Einstellung der Lautstärke, der Möglichkeit der Unterbrechung und/oder der Wiederholung so-

wie bei Graphiken durch die Anpassung der Größe. Die Rückmeldungen können je nach Bedarf beispielsweise hinsichtlich einer Stärken-/Schwächenorientierung, des fokussierten Themengebiets, des Detaillierungsgrades und/oder der gewünschten Bezugsnorm ausgegeben werden. Darüber hinaus können Nutzer-/Nutzerinnenprofile mit persönlichen Angaben erstellt werden, in denen der Lernverlauf abgebildet ist.

Aus der Darstellung zur Mediananreicherung ist weder zu schlussfolgern, dass eine möglichst große Varianz immer sinnvoll ist; noch ist daraus abzuleiten, dass die Ausprägung der Innovation umso stärker ist, je größer die Varianz ist. Die Codierung des Inhalts und die Medieneinbindung sind immer unter Berücksichtigung der Aufgabe und der Zielgruppe zu wählen. Medienreichtum darf nicht zu einer Überforderung der Rezipienten führen. Wie auch in der CLT und der CTML (vgl. Kap. 2.4) beschrieben ist zu vermuten, dass eine zu reichhaltige Darstellung, die mehrere Informationskanäle (z. B. Text, Bild und Ton) beansprucht, so viele kognitive Ressourcen zur Decodierung bindet, dass die eigentliche Lösung der Aufgabe erschwert wird und somit die Konstruktvalidität verschlechtert werden würde. Besonders bezüglich der Zielgruppe der funktionalen Analphabeten und Analphabetinnen ist ein geringer Toleranzbereich zu vermuten (vgl. Kap. 2.3). Die Medienwahl ist an deren Lese- und Schreibkompetenz sowie deren ICT-Literacy anzupassen. So sind die auditive Unterstützung von Instruktionen und die Veranschaulichung von Bildern sinnvoll. Voraussichtlich würden aber zu viele Variationsmöglichkeiten und eine dichte Darstellung von unterschiedlichen Inhalten zu einer *Overcomplication* (Schwabe, 2001, S. 55) führen. Im Idealfall führt der Einsatz von Medien allerdings zu einer Reduzierung von kognitiven Ressourcen, da mehrere Sinneskanäle angesprochen werden und/oder literale Defizite durch auditive Unterstützung kompensiert werden können.

Zusammenfassend liegen die Vorteile einer reichen Medieneinbindung zum einen darin, dass die Herstellung einer oft gewünschten Authentizität und Komplexität nur durch die Mediananreicherung möglich ist. Zum anderen beinhaltet ein hoher Medienreichtum das Potenzial, die Testmotivation aufrecht zu erhalten. Handelt es sich allerdings um rein dekorative Medien, die womöglich nur einen geringen Bezug zum Aufgabeninhalt herstellen, ist mit einer Verschlechterung der Inhaltsvalidität zu rechnen (Laitusis, 2010; Mayer, 2005). Schlussfolgernd ist ein Itemformat bezüglich der Mediananreicherung dahingehend innovativ, wenn das Potenzial neuer Technologien genutzt wird, indem die Medien und deren Kombination im Sinne des Messgegenstands und der Zielgruppe sinnvoll gewählt werden.

Aufgrund dessen, dass sich die Medienanreicherung auf die technische Integration verschiedener Medienarten bezieht, ist diese Dimension dem Organisationsschema Computertechnologie zuzuordnen.

5. Interaktivität: Interaktivität bezeichnet das Aufeinanderbeziehen von wechselseitigen Kommunikations- und Handlungsprozessen von zwei oder mehr aktiven Agenten, die alternierend zuhören, denken und/oder sprechen (Crawford, 2004, S. 29). Dabei werden folgende Aspekte unterschieden: (a) die Schnelligkeit der Reaktion auf die Aktion des Interaktionspartners (z. B. der Klick auf eine Schaltfläche in einem Programm); (b) die Tiefe (beispielsweise der Rechenaufwand eines Schachprogrammes zur Berechnung des nächsten Zuges oder das Handlungsmuster eines computergesteuerten Spielpartners in einem Computerspiel) und (c) die Wahlmöglichkeiten, insbesondere die funktionale Signifikanz einer Entscheidung auf den weiteren Verlauf der Interaktion und die wahrgenommene Vollständigkeit der Wahlmöglichkeiten.

Voraussetzung für die Interaktivität ist, dass die computerbasierte Diagnostik die Handlungen des „Interaktionspartners“ – in diesem Fall der Mensch – eindeutig interpretiert werden, um den Verlauf einer sinnvollen Interaktionskette zu ermöglichen. Demnach sind alle Computerprogramme mit implementierten Feedbackprozessen als interaktiv zu bezeichnen: In Abhängigkeit der Aktionen – der Eingaben – erfolgt eine Rückmeldung des Computers bzw. des Programms. Die Dimension Interaktivität bezieht sich insbesondere auf die programmier-technische Implementation und ist daher ebenfalls dem Organisationsschema Computertechnologie zuzuordnen.

2.2.6.2 Taxonomie innovativer Itemformate

Aus dem vorherigen Kap. geht hervor, hinsichtlich welcher Dimensionen die Itemformate charakterisierbar sind. Eine zu den fünf Dimensionen *Offenheit des Antwortformats*, *Auswertungsalgorithmus*, *Authentizität*, *Medienanreicherung* und *Interaktivität* quer liegende Dimension ist die *Innovativität*. In Abhängigkeit der Dimensionsausprägung verändert sich auch das *innovationspotenzial*. Die folgende Tabelle zeigt eine Taxonomie von innovativen Itemformaten, in der neben den fünf Dimensionen sowohl beide Organisationsschema als auch die Dimension der Innovativität berücksichtigt sind.


Organisationschema	Dimension	Innovationspotenzial		
		niedrig		hoch
Messung	1. Offenheit des Antwortformats	geschlossen	halboffen	offen
	2. Auswertungsalgorithmus Automatisierung Komplexität	nicht automatisiert dichotom	teilautomatisiert polytom	automatisiert komplex
	3. Authentizität	abstrakt	teilkontextualisiert	kontextualisiert
Computertechnologie	4. Medienanreicherung Inhaltsdarstellung Medieneinbindung Personalisierung	monocodal keine Varianz nicht möglich	duocodal wenig Varianz teilweise möglich	multicodal viel Varianz möglich
	5. Interaktivität	niedrig	mittel	hoch

Tabelle 2: Taxonomie innovativer Itemformate (eigene Darstellung in Anlehnung an Wolf, Koppel & Schwedes 2011)

In der linken Spalte sind die Organisationsschemata aufgeführt, denen in der zweiten Spalte die fünf Dimensionen mit ihren Subkategorien zugeordnet sind. Die Dimensionen werden jeweils in ihren extremen und mittleren Ausprägungen qualifiziert. Je stärker ein Item in seiner Ausprägungstendenz zu den in der rechten Spalte aufgeführten Attributen tendiert, als desto größer ist auch das Innovationspotenzial einzustufen.

Der oben aufgeführten Definition von innovativen Itemformaten folgend, bieten offene Antwortformate Möglichkeiten der Messung, die mit geschlossenen Antwortformaten nicht oder nur bedingt möglich sind (z. B. komplexe Problemlösefähigkeit). Die Möglichkeiten zur Abbildung von Komplexität und automatisierter Auswertung ist in konventionellen papierbasierten Tests ebenfalls weniger möglich als in technologiebasierten Umgebungen. Damit einher geht auch die Medienanreicherung und die Interaktivität. Eine multicodale Darstellung von Inhalten mit einer hohen Medienvarianz und Option zur Personalisierung kann die Kriterien eines innovativen Items (Nutzung der Technik, Visualisierungsmöglichkeiten und Erweiterung des Messgegenstands) in einem weitaus höheren Maße erfüllen als es papierbasierte Tests oder einfache Single Choice-Aufgaben am Computer können. Der in der Taxonomie verwendete Begriff *Innovationspotenzial* vermittelt, dass kein unabdingbarer kausaler Zusammenhang zwischen dem Ausprägungsgrad der fünf Dimensionen und der Innovativität

besteht: Eine hohe Medieneinbindung und Authentizität bedingt nicht eine gleichzeitig hohe Innovativität. Im Vordergrund steht immer, die Dimensionen und deren Ausprägungsgrad im Sinne des Messgegenstands und der Zielgruppe zu wählen und miteinander zu kombinieren.

Anhand des folgenden Beispiels wird die Einsatzmöglichkeit der Taxonomie verdeutlicht: Für die Entwicklung einer computerbasierten Diagnostik zur Messung der Schreibfähigkeit bei wenig literalisierten Personen kann die Taxonomie herangezogen werden, um sowohl den Messgegenstand und die -absicht als auch zielgruppenspezifische Charakteristika zu berücksichtigen. Die Schreibfähigkeit kann z. B. über die Rechtschreibfähigkeit von einzelnen Wörtern geprüft werden. Dafür wäre ein offenes Antwortformat (Textfeld oder Freitexteingabe) notwendig. Die Auswertung würde dichotom (richtig/falsch) erfolgen. Eine starke Ausprägung der Authentizität wäre nicht notwendig, da die Rechtschreibung von Wörtern kontextunabhängig messbar ist. Die Misserfolgsbefürchtungen können bei funktionalen Analphabeten und Analphabetinnen hoch sein (vgl. Kap. 2.3), daher ist es unter Umständen hilfreich, authentische Formate einzubeziehen, um die Motivation aufrecht zu erhalten bzw. zu steigern. Für eine Inhaltsdarstellung kann eine duocodale Ausprägung sinnvoll sein, um das Arbeitsgedächtnis zu entlasten und nicht nur Text zu präsentieren, dessen Decodierung, je nach Niveau, bei wenig literalisierten Personen einen enormen Arbeitsaufwand bedeuten kann (Ermüdung und Frustration können die Folge sein). Hinsichtlich der Medienanreicherung ist mindestens eine geringe Varianz sinnvoll, um die Nutzer und Nutzerinnen mit Audiodateien zu unterstützen und die Instruktionen vorlesen lassen zu können. Eine ausgeprägte Personalisierung wäre nicht notwendig; zudem könnte eine zu starke Personalisierung zu einer Überforderung aufgrund der tendenziell niedrigen ICT-Literacy führen. Eine Personalisierung wäre diesbezüglich allerdings gewinnbringend realisierbar hinsichtlich der Abbildung des individuellen Lernverlaufs, indem Lernende einen Fokus setzen und z. B. auswählen können, welchen Kompetenzbereich sie sich anzeigen lassen möchten. Auch die Interaktivität sollte aufgrund der tendenziell niedrigen ICT-Literacy gering gehalten werden. Personen mit einer sehr niedrigen ICT-Literacy können durch hoch interaktive Itemformate leicht überfordert werden.

Zusammenfassend kann diese Taxonomie sowohl für die Entwicklung von Itemformaten und die sinnvolle Abstimmung des Ausprägungsgrades als auch für die Kategorisierung und Bewertung von Itemformaten herangezogen werden. Bei der Entwicklung von Items wird meist von dem zu messenden Konstrukt ausgegangen, wodurch bereits in vielen Fällen eine Präferenz für die Formatauswahl besteht. Die Wahl für die Offenheit des Antwortformats beeinflusst u. U. den Ausprägungsgrad weiterer Dimensionen. Hinsichtlich der personalen

Ressourcen ist die Frage zu stellen, ob die Auswertung automatisiert stattfinden soll oder ob die Auswertung der Aufgaben durch Personen durchgeführt werden kann. Wie erwähnt, können geschlossene Formate im Vergleich zu offenen Formaten leicht automatisiert ausgewertet werden; ein Aufsatz kann nur bedingt automatisiert ausgewertet werden und es bedarf eines hohen Entwicklungs- und Programmieraufwandes (vgl. Kap. 2.2.5 und z.B. Sukkarieh & Stoyanchev, 2009). Vor dem Hintergrund dieser Entscheidungen bzw. Rahmenbedingungen ist bzgl. der prognostischen Validität und der Testmotivation weiterhin zu klären, inwieweit die zu entwickelnden Items authentisch, interaktiv und medienreich sein sollten (Wolf u. a., 2011, S. 139). Dafür ist insbesondere die Zielgruppe zu berücksichtigen. Mit der Taxonomie können die Ausprägungen der Items unter Berücksichtigung der Messabsicht, des Messgegenstands und der Zielgruppe gewählt werden, um eine sinnvolle Umsetzung zu ermöglichen.

2.3 Zielgruppenspezifische Voraussetzungen

Wie bereits in den vorherigen Kapiteln angedeutet, sind Diagnoseverfahren immer kontextgebunden. Das bedeutet, die Verfahren sind an die Zielgruppe und die Rahmenbedingungen anzupassen, indem zielgruppenspezifische Charakteristika und der potenzielle Einsatzbereich berücksichtigt werden. Um ein computerbasiertes Diagnoseinstrument für funktionale Analphabeten und Analphabetinnen entsprechend dieser Zielgruppe entwickeln und einsetzen zu können, sind folgende Kontextvariablen zu berücksichtigen: Literalität, Struktur/Verteilung des Ausmaßes funktionalen Analphabetismus⁷, Herausforderungen in der Alphabetisierungspraxis sowie die Computerkompetenzen von funktionalen Analphabeten und Analphabetinnen. Diese werden im Folgenden dargestellt.

2.3.1 Literalität

Der Begriff Literalität ist auf den lateinischen Begriff *littera* für *Buchstabe* zurückzuführen, was vorerst nur auf die Fähigkeit, mit Buchstaben umgehen zu können, deutet aber keineswegs eine kompetenzbezogene Definition liefert. Eine Definition von Literalität lieferte die UNESCO 1978:

"A person is literate who can with understanding both read and write a short simple statement on his everyday life" (UNESCO, 1978).

Eine weitere Definition wurde von der OECD formuliert: Literalität ist "die Verwendung von gedruckten und geschriebenen Informationen, um in der Gesellschaft zurechtzukommen, eigene Ziele zu erreichen und eigenes Wissen sowie die individuellen Möglichkeiten zu entwickeln" (OECD, 1995, S. 16). Diese

Definition wurde zudem in der IALS-Studie verwendet. In vorangegangenen Diskussionen wurde der Begriff Literalität zunehmend mit Grundbildung synonym verwendet (Grotlüschen, 2011, S. 14). Begründet ist dies womöglich in der Übersetzung von *literacy* im Englischen in *Literalität* ODER *Grundbildung* im Deutschen. So wird auch im deutschen Sprachraum beispielsweise von *computer-* bzw. *ICT-Literacy* gesprochen. Im englischsprachigen Raum finden vor allem auch Begriffe wie *political literacy* Verwendung (Linde, 2007, S. 94).

Deutlich wird, dass in beiden Definitionen (der UNESCO und der OECD) eine Kompetenz- bzw. Fähigkeitsdimension verwendet wird, nämlich "verstehen", "zurechtkommen" und „Entwicklung“. Eine weitere Dimension ist die Gesellschaftliche mit den Beschreibungen "everyday life" und "in der Gesellschaft zurechtkommen". Die gesellschaftliche Dimension impliziert, dass es sich bei diesem Begriff (wie auch beim Begriff "Funktionaler Analphabetismus" - vgl. folgendes Kap.) um ein relationales Konzept handelt. Was Nickel (2011) in Bezug auf Funktionalen Analphabetismus schreibt, kann hier auch auf den Begriff Literalität bezogen werden, "da die Anforderungen an Schriftsprachlichkeit historisch und kulturell wandelbare Größen sind und daher variieren" (Nickel, 2011, S. 54). Internationale Literalitätsstandards (hier verstanden als Lese- und Schreibfähigkeiten) können somit nur anhand "kulturell-funktionaler" Dimensionen definiert werden, die wiederum in Abhängigkeit des gesellschaftlichen Kontextes entsprechend betrachtet werden müssen.

Abhängig vom gesellschaftlichen Kontext wird dementsprechend zwischen literalisierten und nicht bzw. wenig literalisierten Personen unterschieden. Der zuletzt genannten Gruppe gehören die Analphabeten/Analphabetinnen sowie die funktionalen Analphabeten und Analphabetinnen an. Da die funktionalen Analphabeten und Analphabetinnen die Zielgruppe dieser Arbeit darstellen, wird der funktionale Analphabetismus im Folgenden näher erläutert.

2.3.2 Funktionaler Analphabetismus

Funktionaler Analphabetismus ist ein weltweit existierendes Phänomen. Auch in hoch industrialisierten Ländern und Ländern mit einem hohen Bildungsstand tritt das Phänomen auf. Funktionaler Analphabetismus ist ein Begriff, der sich auf die Lese- und Schreibkompetenzen von Erwachsenen bezieht. Eine Definition, die nicht an operationalisierten Sprachstandsbeschreibungen festhält kann somit nur auf einer höheren Ebene erfolgen.

In einem Report zur Grundbildung in Europa wird von der *French National Agency for the Fight against Illiteracy* festgehalten:

"Literacy and numeracy definitions need to be operational rather than global or static and are highly dependent on the actual assessment tool. The com-

plexity of linking definition to literacy level need always be borne in mind and the purposes of those used in testing be quite clear from the earliest stages through to actual use of results" (ANLCI - French National Agency for the Fight against Illiteracy, 2009, S. 7).

Sprachstandserhebungen geben somit immer Werte an, die auf das jeweilige Erhebungsinstrument und die erhobene Sprache referieren und mit dem dahinterliegenden Verständnis von Literalität verknüpft sind. Dennoch bestehen Bemühungen, eine möglichst weitreichende Definition zu etablieren. So hat eine Gruppe von Wissenschaftlern und Wissenschaftlerinnen aus dem Förderschwerpunkt „Forschung und Entwicklung zur Alphabetisierung und Grundbildung Erwachsener“ (Laufzeit 2007-2011) im Jahr 2010 eine Definitionsgrundlage zum Funktionalen Analphabetismus geliefert²⁹:

„ „Funktionaler Analphabetismus“ ist gegeben, wenn die schriftsprachlichen Kompetenzen von Erwachsenen niedriger sind als diejenigen, die minimal erforderlich sind und als selbstverständlich vorausgesetzt werden, um den jeweiligen gesellschaftlichen Anforderungen gerecht zu werden. Diese schriftsprachlichen Kompetenzen werden als notwendig erachtet, um gesellschaftliche Teilhabe und die Realisierung individueller Verwirklichungschancen zu eröffnen.

Unter schriftsprachlicher (literal)er Kompetenz ist die Fähigkeit zu verstehen, sich der Schrift als Kommunikationsmittel zu bedienen. [...] Dies ist gegenwärtig zu erwarten, wenn eine Person nicht in der Lage ist, aus einem einfachen Text eine oder mehrere direkt enthaltene Informationen sinnerfassend zu lesen³⁰ und/oder sich beim Schreiben auf einem vergleichbaren Kompetenzniveau befindet.“

²⁹ Die Definition ist ohne Angabe von Autoren oder Autorinnen auf www.grundbildung.de/daten/Grundlagen/definition (zuletzt geprüft am 03.11. 2014) erschienen.

³⁰ Dies entspricht in der PISA-Studie der Kompetenzstufe 1 („Eine oder mehrere unabhängige, leicht auffindbare Information(en) lokalisieren; Voraussetzung für das Auffinden der Information: wenig konkurrierende Informationen im Text“; vgl. PISA 2000, Opladen 2001, S. 89). In der IALS-Studie

Ihren Ausgangspunkt nimmt diese Definition in der Definition der UNESCO von 1978 und folgt einer Art Baukastenprinzip, indem anfangs das Problem benannt und im weiteren Verlauf näher bestimmt wird. Nach Angabe der Verfassenden können je nach Fragestellung oder Projektkontext die angegebenen Erläuterungen variiert oder ergänzt werden. Auch bei dieser Definition wird sich wie bei der Definition von Literalität einer kulturell-funktionalen Dimension bedient. Die leo.- Level-One-Studie liefert einen Beitrag hin zu einer stärkeren Operationalisierung, indem ein Bezug zu den Kompetenz-Level-beschreibungen (vgl. Kap. 4.1 zum lea.-Projekt) hergestellt wird:

„Die UNESCO spricht von funktionalem Analphabetismus bei Unterschreiten der vollen Teilhabe im Lesen, Schreiben und Rechnen. [...] leo. operationalisiert dies als Unterschreiten des Alpha-Level 3. leo. differenziert [...] nach Alpha-Level 4, 5 und 6 aus und rechnet das Unterschreiten des Alpha-Level 4 dem funktionalen Analphabetismus zu“ (Grotlüschen, Riekmann & Buddeberg, 2012, S. 18).

Die leo.-Definition des funktionalen Analphabetismus ist somit das Unterschreiten der Textebene, was dem Unterschreiten des Alpha-Level 4 entspricht (Grotlüschen, Riekmann & Buddeberg, 2012, S. 18).

Eine weltweit einheitliche hinreichend operationalisierte Definition existiert jedoch nicht (Grotlüschen & Riekmann, 2012). Die Schwierigkeit einer einheitlichen Definition liegt möglicherweise darin, dass der Sprache und der Sprachfähigkeit linguistische Sprachspezifika zugrunde liegen. Daher erscheint der Weg einer Definition über die gesellschaftliche Teilhabe im Sinne der Vergleichbarkeit sinnvoll. Doch sind nationale Studien verschiedener Länder, in denen das Ausmaß des Funktionalen Analphabetismus angegeben wird, spätestens auf der Ebene der Operationalisierung und der sprachspezifischen Kompetenzmodelle nur schwer oder nur unter Vorbehalt vergleichbar. Daher werden die Größenordnungen des Funktionalen Analphabetismus in Deutschland, Frankreich und Großbritannien im Folgenden dargestellt, ohne einen direkten statistischen Bezug zwischen den Ländern herzustellen. Der Fokus liegt dabei auf der Größenordnung in Deutschland.

(„International Adult Literacy Survey“; vgl. OECD/Statistics Canada 1995) entspricht dies ebenfalls der Kompetenzstufe I des Leseverständnis bei Prosa-Texten.

2.3.3 *Funktionaler Analphabetismus in Deutschland, England und Frankreich*

Im Jahr 2010 wurde deutschlandweit die leo.-Level-One-Studie³¹ durchgeführt. Die leo.-Studie hatte das Ziel, ein Benchmark zur Größenordnung des funktionalen Analphabetismus der deutsch sprechenden Erwachsenen zu definieren. Dabei baut die leo.-Studie auf Vorarbeiten des lea.-Projektes (blogs.epb.uni-hamburg.de/lea) auf. Im lea.-Projekt wurden die Theoriedifferenzierung, die Levelbeschreibungen und die Itementwicklung vorbereitet. Die Studie hatte nicht das Ziel, die Literalität der gesamten Bevölkerung zu erheben, sondern zu ermitteln, wie viele dem niedrigsten Kompetenzbereich – dem sog. Level One – zuzuordnen sind und diesen auszudifferenzieren.

14,5% und somit 7,5 Millionen der erwerbsfähigen erwachsenen Bevölkerung (18-64 Jahre) sind in Deutschland vom funktionalen Analphabetismus betroffen. Wie bereits oben erwähnt bedeutet dies, dass diese Personen die Textebene unterschreiten. Eine Person kann somit zwar einzelne Sätze lesen oder schreiben, nicht jedoch zusammenhängende Texte. Diese Personen sind aufgrund ihrer begrenzten schriftsprachlichen Kompetenzen nicht in der Lage, am gesellschaftlichen Leben in angemessener Form teilzuhaben (Grotlüschen u. a., 2012, S. 20).

Weitere 25,9% (13,3 Millionen) können nur fehlerhaft schreiben. Diese Personen können trotz gebräuchlicher Wörter nur langsam lesen und/oder fehlerhaft schreiben und beherrschen nicht die Rechtschreibung, wie sie bis zum Ende der Grundschule unterrichtet wird. Kumuliert bedeutet dies, dass ca. 40% ihre literalen Kompetenzen noch deutlich verbessern können (Grotlüschen u. a., 2012, S. 20).

Überraschend erscheinen die Ergebnisse zum funktionalen Analphabetismus und der Erwerbstätigkeit: Fast 57% der funktionalen Analphabeten und Analphabetinnen in Deutschland geben an, erwerbstätig zu sein (Grotlüschen, 2012, S. 137).

³¹ blogs.epb.uni-hamburg.de/leo (zuletzt geprüft am 03.11. 2014)

Anteil	Funktionaler Analphabetismus			Summe	Fehlerhaftes Schreiben		Bevölkerung gesamt
Alpha-Level	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 1 - \alpha 2$	$\alpha 4$	$> \alpha 4$	
Erwerbstätig	54,8%	54,2%	58,0%	56,9%	64,5%	69,5%	66,4%
Arbeitslos	19,6%	21,6%	14,7%	16,7%	8,9%	4,8%	7,6%
Erwerbsunfähig	2,7%	2,3%	2,3%	2,3%	1,5%	0,9%	1,3%
Hausfrau/-mann, Elternzeit	17,45	10,8%	9,4%	10,1%	8,2%	7,9%	8,3%
Rentner/Rentnerin	5,1%	6,3%	6,4%	6,3%	6,2%	3,8%	4,8%
In Ausbildung	0,4%	4,0%	7,9%	6,5%	9,9%	11,6%	10,4%
Sonstiges	0,0%	0,8%	1,4%	1,2%	0,9%	1,4%	1,2%
Summe	100%	100%	100%	100%	100%	100%	100%

Tabelle 3: Beruflicher Status funktionaler Analphabetismus und fehlerhaftes Schreiben (Grotlüschen, 2012, S. 141)

Wie aus der Tabelle 3 hervorgeht, sind fast 55% der Personen auf Alpha-level 1 erwerbstätig, wobei das Alpha-Level 1 bedeutet, dass die zugehörigen Personen die Wortebene beim Lesen und Schreiben nicht erreichen. Von den Personen, die dem Alpha-Level 2 zuzuordnen sind, gehen 54,2% einer Erwerbstätigen Beschäftigung nach. Diese Personen erreichen nicht die Satzebene. Von den Zugehörigen des Alpha-Level 3 sind 58% erwerbstätig, wobei diese Personen kurze Sätze lesen und schreiben können (Grotlüschen & Riekmann, 2011, S. 4). Insgesamt sind somit deutlich mehr funktionale Analphabeten und Analphabetinnen erwerbstätig (56,9%) als arbeitslos (16,7%). Genauere Analysen zeigen, dass 19,7% der funktionalen Analphabeten und Analphabetinnen einer geringfügigen Beschäftigung nachgehen und 86,1% wöchentlich 35 Stunden oder mehr arbeiten. Im Vergleich zur Gesamtbevölkerung sind funktionale Analphabeten und Analphabetinnen überproportional in prekären Beschäftigungsverhältnissen. Allerdings zeichnen sich die prekären Beschäftigungsverhältnisse nicht durch die Befristung von Arbeitsverträgen, sondern in der geringen Entlohnung ab (Grotlüschen, 2012, S. 163).

In England wird zwischen drei Entry Level unterschieden, wobei diese zusammen den IALS Level One abbilden (Grotlüschen & Riekmann, 2012, S. 21). So wird in England von 14,9% funktionaler Analphabeten und Analphabetinnen zwischen 16 und 65 ausgegangen (Department for Business, Innovation and Skills, 2011, S. 5).

In Frankreich wurde mit dem „Information et Vie Quotidienne“ 2004-2005 eine Anzahl von 3,1 Millionen funktionalen Analphabeten und Analphabetinnen erhoben, was neun Prozent der erwachsenen Bevölkerung entspricht. Dabei sind allerdings ausschließlich diejenigen befragt worden, die in Frankreich die Schule besucht haben (ANLCI - French National Agency for the Fight against Illiteracy, 2009).

In Frankreich ist der Term "funktionaler Analphabetismus" allerdings nicht gebräuchlich. In Frankreich wird zwischen drei Formen mangelnder Literalität unterschieden: Personen, die zur Schule gegangen sind, aber Lesen und Schreiben nicht weiter praktiziert haben, Personen, die nie lesen und schreiben gelernt haben sowie Personen, die nicht in Frankreich literarisiert sind und somit nicht in der Lage waren, ihre Lese- und Schreibkompetenzen auszudrücken.

Für die Literalität³² gibt die Studie IALS für Deutschland eine Anzahl von 14,4%, in England von 21,8% auf dem Level One für Literalität aus. Frankreich hatte sich aus der Erhebung zurückgezogen nachdem sich abzeichnete, dass der Wert nahe der Vierzigprozentmarke liegen könnte (Grotlüschen u. a., 2012, S. 23).

Neue Ergebnisse liefert die PIAAC-Studie. In dieser wird zwar nicht die Literalität hinsichtlich funktionalem Analphabetismus ausgewiesen, doch ist die Angabe des Leseverständnisses im Vergleich zu den OECD-Ländern einen Hinweis, wie stark funktionaler Analphabetismus im Vergleich beispielsweise zu Deutschland ausgeprägt sein mag: Frankreich rangiert signifikant deutlich im Bereich Lesekompetenz unter dem OECD-Durchschnitt mit einem Mittelwert von 262 Punkten, Deutschland liegt ebenfalls signifikant unter dem Durchschnitt, weist aber einen höheren Mittelwert von 270 Punkten auf. 18% der deutschen Bevölkerung kommen über die niedrigste Kompetenzstufe von PIAAC nicht hinaus. Spitzenreiter ist Japan mit 296 und Schlusslicht ist Italien mit 250 Punkten (Rammstedt u. a., 2013, S. 13).

Ging man vor der leo.-Studie noch von vier Millionen funktionalen Analphabeten und Analphabetinnen aus, zeigen die leo.- und die PIAAC-Studien doch ein erheblich größeres Ausmaß des funktionalen Analphabetismus. Aller-

³² Im englischen Sprachraum wird von *prose literacy* gesprochen. Unter Prosa-Literalität ist die Literalität zu verstehen, die im Alltag genutzt wird (beispielsweise das Verstehen von Nachrichten). In diesem Kontext ist die Prosa-Literalität gleichbedeutend mit dem Begriff der Literalität im deutschen Sprachraum. Daher wird hier auf den Begriff der Literalität zurückgegriffen.

dings ist anzumerken, dass die Zahl vier Millionen eine Schätzung war, in der nur deutsche Muttersprachler und Muttersprachlerinnen berücksichtigt wurden. In der leo.-Studie wurden nicht nur die in Deutschland literalisierten Personen, sondern auch Personen mit Migrationshintergrund einbezogen. Voraussetzung für die Teilnahme war eine vorgelagerte mündliche Überprüfung, ob die Deutschkenntnisse ausreichen, die Aufgaben verstehen zu können. Somit sind die geschätzte Zahl von vier Millionen und die erhobene Zahl von 7,5 Millionen vor dem Hintergrund der befragungsspezifischen Voraussetzungen nicht als eine Verdopplung oder ein enormer Anstieg des funktionalen Analphabetismus zu betrachten. Zusammenfassend weisen die aktuellen Studien auf einen enormen Alphabetisierungsbedarf hin. Demgegenüber steht die Alphabetisierungspraxis allerdings vor diversen Herausforderungen.

2.3.4 Herausforderungen in der Alphabetisierung in Deutschland

Trotz 7,5 Millionen funktionalen Analphabeten und Analphabetinnen nehmen in Deutschland nur ca. 30.000 Personen an Alphabetisierungskursen teil (Huntemann & Reichart, 2013, S. 30). Ursachen werden sowohl seitens der Betroffenen selbst als auch aus der Sicht von Weiterbildungsanbietern gesehen. Aus der Perspektive von funktionalen Analphabeten und Analphabetinnen werden für die geringe Teilnahme folgende Gründe genannt (Koppel & Wolf, 2014; Steuten, 2013)³³:

- a) geringes Angebot (ca. jede dritte Volkshochschule bietet Alphabetisierungskurse an (Rosenblatt & Lehmann, 2013b, S. 57);
- b) niedriger Bekanntheitsgrad;
- c) Vermeidung, empfundene und/oder tatsächliche Schwächen offen zu legen, beispielsweise auf Grund von Schamgefühlen (vgl. z.B. Döbert u. a., 2000; Egloff, 1997; Füssenich, 2004; Schladebach, 2007);
- d) Subjektiv wird kein Bedarf an Alphabetisierung empfunden.

In den Alphabetisierungskursen selbst sehen sich sowohl Kursleitende als auch Lernende mit weiteren Herausforderungen konfrontiert:

- a) wenig standardisierte Curricula (erst mit der Einbindung der Alpha-Level in das Rahmen-Curriculum der Volkshochschulen im Jahr 2014 wurde eine Grundlage für einheitliche und auf empirisch überprüften Standards basierende Curricula erreicht);
- b) keine einheitlichen Lehr-/Lernmethoden;
- c) keine klare Vorstellung über die Abschlussziele (Rosenblatt & Lehmann, 2013b) (mit dem neuen Rahmencurriculum des Deutschen

³³ Die Herausforderungen in der Alphabetisierungen wurden bereits in Koppel & Wolf (2014) erläutert.

Volkshochschulverbandes ist allerdings die Basis für die Formulierung einheitlicher Ziele geschaffen);

- d) die Nutzung diagnostischer Verfahren ist umstritten (Bonna & Nienkämper, 2011);
- e) Befragungen und Tests stoßen vielfach auf Abwehrreaktionen (Schladebach, 2007), da diese meist eine Defizitorientierung aufweisen.

Die Teilnahme an Alphabetisierungskursen führt somit nur zu einem begrenzten Lernerfolg: Die Lernfortschritte reichen in den meisten Fällen nicht aus, um bei der Mehrzahl der Teilnehmenden ein schriftsprachliches Kompetenzniveau oberhalb des funktionalen Analphabetismus zu erreichen (Rosenblatt & Lehmann, 2013b, S. 72). Aus der Perspektive der Kursleitenden erscheint zudem problematisch, dass wenig erwachsenengerechte Materialien zur Diagnose und Förderung von funktionalen Analphabeten und Analphabetinnen vorhanden sind (Heinemann, 2011). Zudem ist eine individuelle Diagnose des Lernstands meist zeitaufwendig. Auch müsse berücksichtigt werden, dass es um Personen gehe, die häufig Lernbehinderungen aufweisen und in schwierigen sozialen Verhältnissen leben (Rosenblatt & Lehmann, 2013b, S. 57).

Für die Alphabetisierungsarbeit wird daher gefordert, Grundbildungsangebote auf eine tragfähige Basis mit erhöhter Professionalität zu stellen (Frieling & Rustemeyer, 2011, S. 39) sowie Erfolgskriterien und Lernziele zu formulieren (Rosenblatt & Lehmann, 2013b, S. 73).

Als Konsequenzen werden eine Ausweitung und Diversifizierung des bisher relativ schwachen Angebots gefordert (Rosenblatt & Lehmann, 2013a, S. 10). Es sollen differenzierte Angebote geschaffen werden, die sinnvoll die heterogenen Kompetenzstrukturen der funktionalen Analphabeten und Analphabetinnen aufgreifen und berücksichtigen. Auf der Ebene der Kursarbeit sind daher Möglichkeiten bereitzustellen, mit denen zuerst festgestellt werden kann, welche Kompetenzen Kursteilnehmende bereits erworben haben und welche zu fördern sind (Eingangsdiagnostik). Es bedarf daher keiner (summativen) Diagnostik zu Selektionszwecken (vgl. Kap. 2.1), sondern einer Diagnostik zur differenzierten Abbildung der Kompetenzen, die eine Ableitung gezielter Fördermaßnahmen ermöglicht. Eine Möglichkeit, diesen Herausforderungen zu begegnen ist der Einsatz einer computerbasierten Diagnostik. Beim Einsatz computerbasierter Diagnostik ist allerdings immer auch zu berücksichtigen, wie stark die ICT-Literacy ausgeprägt ist und ob die Computerfähigkeit (bzw. ICT-Literacy) einen Einfluss auf das Antwortverhalten hat, um die computerbasierte Diagnostik an die Voraussetzungen der potenziellen Nutzer und Nutzerinnen anpassen zu können.

2.3.5 Informations- und Computertechnologie (ICT)-Literacy

ICT-Literacy wird im Kontext der PISA-Studie folgendermaßen definiert:

„ICT literacy is the interest, attitude, and ability of individuals to appropriately use socio-cultural tools, including digital technology and communication tools, to access, manage, integrate, and evaluate information, construct new knowledge, and communicate with others in order to participate effectively in society“ (Schleicher, 2008, S. 6).

Nach der OECD ist die ICT-Literacy die Voraussetzung bzw. das Werkzeug für effizientes und selbstgesteuertes Lernen mit Informations- und Kommunikationstechnologien (Schäffer, 2006). Diese Definition ist logischerweise an die der Literacy angelegt „the ability to understand and employ printed information in daily activities, at home, at work and in the community – to achieve one’s goals, and to develop one’s knowledge“ (OECD, 2000, S. X [sic!]). Unter Berücksichtigung der Technologieentwicklung wird der Rückbezug auf dieses Konzept in Zukunft nicht ausreichend sein, da Informationen auch auditiv, visuell und wahrscheinlich bald auch haptisch und olfaktorisch dargeboten werden (vgl. auch zur Kritik des ICT-Literacy-Konzepts Schäffer, 2006; Sting, 2005).

Aus länderübergreifenden Studien geht hervor, dass es zwischen den beteiligten Ländern große Unterschiede bezüglich der Computer- und Internetnutzung gibt. So besteht ein signifikanter Zusammenhang zwischen der Höhe des Einkommens, dem formalen Bildungsstand und dem ausgeübten Beruf auf der einen sowie den Literacy-Skills und dem Zugang zum Computer auf der anderen Seite (OECD, 2005, S. 192). Dieses Phänomen wird auch als „digital divide“ beschrieben (vgl. z. B. Montagnier, 2011). Allerdings ist ein zunehmender Anstieg hin zu einer „Vollausstattung“ aller Haushalte mit Computern und Internetzugang in Deutschland zu verzeichnen. Waren im Jahr 2006 noch 64% der Haushalte mit Computern ausgestattet, so waren es 2012 bereits 81% (Statistisches Bundesamt, 2012, S. 27). Der mindestens gelegentliche Internetzugang stieg von 59,5% im Jahr 2006 auf 79,1% im Jahr 2014. Die tägliche Internetnutzung liegt 2014 bei 58,3% (ARD/ZDF-Medienkommission, 2014).

Vom Statistischen Bundesamt werden in regelmäßigen Abständen Erhebungen zu Privaten Haushalten herausgegeben und in Fachserien veröffentlicht. In der Fachserie 15 wird die Nutzung von Informations- und Kommunikationstechnologien dargestellt (Statistisches Bundesamt, 2012). Demnach sind in Deutschland 81% der Haushalte mit Computern ausgestattet, 79% besitzen einen

Internetzugang. In der folgenden Tabelle sind Ergebnisse, aufgeschlüsselt nach Bildungsstand³⁴, aufgeführt:

Nutzung Bildungs- stand	Nutzung Handy/ Smart- phone	Compu- ternutzung jeden/fast jeden Tag	Internetnut- zung inner- halb der letzten drei Monate	Senden/ Empfan- gen von Emails	Mitteilungen in soziale Netzwor- ken, Foren, Blogs oder Chaträumen einstellen
niedrig	76	77	61	88	59
mittel	86	80	77	92	38
hoch	90	88	87	96	34

Tabelle 4: Ausstattung der Haushalte mit Computern 2012 (Statistisches Bundesamt 2012, S.17)

Fast jeden Tag den Computer nutzen 77% mit einem niedrigen Bildungsstand, mit einem mittleren Bildungsstand 80% und mit einem hohen Bildungsstand 88% (Statistisches Bundesamt, 2012, S. 12); 76% mit einem niedrigen, 86% mit einem mittleren und 90% mit einem hohen Bildungsabschluss nutzen ein Handy oder Smartphone (Statistisches Bundesamt, 2012, S. 11). 96% der Personen mit hohem und 88% der Personen mit niedrigem Bildungsabschluss senden und empfangen Emails. Ein umgekehrtes Verhältnis zeigt sich bei den Kategorien „Mitteilungen in soziale Netzwerken, Foren, Blogs oder Chaträumen einstellen“ (34 der Personen mit hohem und 59% der Personen mit niedrigem Bildungsabschluss) und „Telefonieren/Videotelefonate“ (33% und 29%) (Statistisches Bundesamt, 2012, S. 17). Ebenfalls deutliche Unterschiede zeigen sich in den Kategorien „Spiele, Bilder, Filme oder Musik (ab-)spielen/herunterladen“ (60%, 44% und 48%), „Netzwerkspiele mit anderen Personen spielen“ (40%, 26%, 15%) sowie das „Hochladen eigener erstellter Texte, Fotos, Videos, Musik usw. auf Webseiten“ (36%, 27%, 28%) (Statistisches Bundesamt, 2012, S. 18).

Bezüglich der Computerkenntnisse steigt die Diskrepanz zwischen dem Bildungsstand: So können beispielsweise 66% mit einem niedrigen Bildungsabschluss „Kopieren und Einfügen“ von Informationen in ein Dokument, wohingegen 85% mit hohem Bildungsabschluss diese Kenntnis besitzen und anwenden

³⁴ Der in den Tabellen aufgeführte Bildungsstand basiert auf den Bildungsstufen nach ISCED (International Standard Classification of Education), der internationalen Standardklassifikation des Bildungswesens. Diese werden auch von der OECD genutzt. Demnach sind dem niedrigen Bildungsstand die Ausbildungsstufen bis zum Abschluss der Hauptschule, der Realschule, des Gymnasiums (Klassen 5-10), Berufsaufbauschule sowie Berufsvorbereitungsjahr zugeordnet. Dem mittleren Bildungsabschluss sind, Gymnasium, Oberschulen, Duale Berufsausbildung sowie Berufsfachschule zugeordnet. Der Kategorie des hohen Bildungsabschlusses gehören Fachhochschule, Universität, Fachschule, Fachakademie, Schulen des Gesundheitswesens sowie Promotion und Habilitation an.

können; „Komprimieren von Daten“ können 31% mit niedrigem und 52% mit hohem Bildungsabschluss (Statistisches Bundesamt, 2012, S. 13).

Für die vorliegende Arbeit relevant sind insbesondere die Kennzahlen der Personen mit niedrigem Bildungsabschluss. Die Zahlen zur Informations- und Kommunikations-Technologie- (IKT)-Nutzung und den IKT-Kenntnissen sind erste Hinweise darauf, wie funktionale Analphabeten und Analphabetinnen diesbezüglich zu charakterisieren sind. So ist festzuhalten, dass Personen mit niedrigem Bildungsabschluss tendenziell weniger erfahren im Umgang mit Computern im Vergleich zu Personen mit hohem Bildungsabschluss sind. Umgekehrt verhält es sich allerdings im Bereich der Internetaktivitäten in den Kategorien „Mitteilungen in soziale Netzwerke, Foren, Blogs oder Chaträumen einstellen“, „Telefonieren/Videotelefonate, Spiele, Bilder, Filme oder Musik (ab-)spielen/herunterladen“, „Netzwerkspiele mit anderen Personen spielen“ sowie das „Hochladen eigener erstellter Texte, Fotos, Videos, Musik usw. auf Webseiten“. Diese werden häufiger von Personen mit niedrigem Bildungsabschluss als von Personen mit einem hohen Bildungsabschluss genutzt.

Diesbezüglich ist anzumerken, dass die Personen mit niedrigem Bildungsstand zwar die funktionalen Analphabeten und Analphabetinnen einschließen, diese aber nur einen kleinen Teil dieser Gruppe ausmachen. Daher ist von den eben vorgestellten Daten der Gruppe mit niedrigem Bildungsstand nicht auf die Gruppe der funktionalen Analphabeten und Analphabetinnen zu schließen (im empirischen Teil dieser Arbeit werden erste Erhebungen zur Computernutzung und -kompetenz von funktionalen Analphabeten und Analphabetinnen vorgestellt). Weitere Informationen zur ICT-Literacy von Personen im Grundbildungsbereich liefert die aktuelle PIAAC-Studie.

In PIAAC 2012 wurde auch die technologiebasierte Problemlösekompetenz erhoben³⁵, die definiert wurde als „Kompetenz, digitale Technologien, Kommunikationshilfen und Netzwerke erfolgreich für die Suche, Vermittlung und Interpretation von Informationen zu nutzen“ (Rammstedt u. a., 2013, S. 4). 44,9% der Bevölkerung befindet sich hinsichtlich der technologiebasierten Problemlösekompetenz nur auf der Stufe I oder niedriger (Zabal u. a., 2013, S. 70), weitere 17,7% hatten entweder keine Computererfahrung, bestanden die IT-(Vor-) Übung des Tests nicht oder verweigerten sich, den Test computerbasiert durchzuführen. Im OECD-Durchschnitt lagen 41,7% auf oder unter Stufe I. Personen ohne Computererfahrung erreichten durchschnittlich nur deutlich niedrigere Werte in den Bereichen Lesekompetenz (227 vs. 276 Punkte) und alltagsmathematische Kompetenzen (213 vs. 282 Punkte) als Personen, bei denen die Kompetenzmessung computergestützt durchgeführt werden konnte (Zabal u. a., 2013,

³⁵ Die folgenden Überlegungen zu den PIAAC-Ergebnissen werden in Wolf und Koppel (2014) veröffentlicht.

S. 68). Dies lässt den Umkehrschluss zu, dass Personen mit niedriger Lesekompetenz auch weniger Computererfahrung haben und offensichtlich ein gewisser Zusammenhang zwischen der Lesekompetenz, der alltagsmathematischen Kompetenzen und der Computererfahrung besteht, was auch im folgenden Zitat zum Ausdruck kommt:

„Damit neue Technologien kompetent und zweckdienlich zur Lösung von alltäglichen Problemen, zum Beispiel zur Informationsbeschaffung, eingesetzt werden können, sind nicht nur die grundlegenden technische Handhabung von Hard- und Software von Bedeutung, sondern insbesondere auch kognitive Fähigkeiten, wie Lese-, mathematische und Problemlösekompetenz wichtig (International ICT Literacy Panel, 2007)“ (Zabal u. a., 2013, S. 60).

Zusammenfassend lassen die Daten die Schlussfolgerung zu, dass funktionale Analphabeten und Analphabetinnen eine tendenziell niedrige ICT-Literacy vorweisen und diese Voraussetzungen in der Entwicklung und Gestaltung von computerbasierter Diagnostik zu berücksichtigen sind.

2.3.6 *Computerbasierte Lern- und Diagnoseinstrumente im Grundbildungsbereich Erwachsener*

Im Grundbildungsbereich Erwachsener sind bisher wenige Instrumente zur Literalitätsmessung und -förderung vorhanden. Insbesondere fehlen Instrumente, die eine kompetenzorientierte Diagnostik anbieten und somit auch das Potenzial für eine förderdiagnostische Tätigkeit besitzen. In Deutschland werden insbesondere zwei computerbasierte Instrumente für den Grundbildungsbereich eingesetzt: das Lernportal „ich-will-lernen.de“³⁶ und das Lernadventure „Winterfest“³⁷. Beide Instrumente wurden vom BMBF gefördert. „ich-will-lernen.de“ ist Deutschlands größtes offenes Lernportal mit mehr als 31.000 kostenlosen Übungen zur Alphabetisierung und Grundbildung. Das Portal eignet sich für Einsteigende als auch Personen mit Vorkenntnissen und kann anonym genutzt werden. Seit dem Jahr 2004 wurden mehr als 389.000 Passwörter vergeben (BMBF, 2014). Das Lernadventure Winterfest ist ein Lernspiel, ebenfalls für Jugendliche und Erwachsene. Aufgaben und Übungen sind in Geschichten eingebettet; Minispiele und Rätsel sollen die Teilnehmenden motivieren, ihre Lese-, Schreib- und Rechenfähigkeiten zu verbessern. Das Spiel kann sowohl privat als auch in der

³⁶ www.ich-will-lernen.de (zuletzt geprüft am 03.11. 2014)

³⁷ www.lernspiel-winterfest.de (zuletzt geprüft am 03.11. 2014)

Kursarbeit eingesetzt werden. Zusätzlich existieren Lehr- und Lernmaterialien, welche ergänzend zum Spiel eingesetzt werden können.

Der Fokus liegt bei diesen Plattformen allerdings nicht auf der Kompetenzmessung, sondern auf dem Lernen von Lesen, Schreiben und Rechnen. Diesen Instrumenten liegt kein validiertes Kompetenzmodell zu Grunde. Somit sind diese für die Diagnostik und insbesondere für die Förderdiagnostik nur bedingt geeignet. Ein Instrument, das förderdiagnostischen Prinzipien genügen soll ist die Online-Testumgebung *otu.lea*. Dessen Entwicklung und Evaluation wird im weiteren Verlauf dieser Arbeit vorgestellt.

Auf internationaler Ebene existieren zahlreiche weitere Online-Plattformen zum Lernen von Lesen, Schreiben und Rechnen. An dieser Stelle soll lediglich einige wenige dieser Programme kurz vorgestellt werden. Die Lernplattform *GCFLearnFree.org*³⁸ bietet Lernmaterialien und Aufgaben im Bereich Technologie (Computerfähigkeiten wie beispielsweise Nutzung von Word und Excel), Lesen (sowohl für Muttersprachler und Muttersprachlerinnen als auch für Personen, die Englisch als Zweitsprache lernen) und Mathematik. Insgesamt stellt die *GCFLearnFree.org* über 85 Übungsthemen und 750 verschiedene Übungseinheiten kostenlos zur Verfügung, die bisher von mehreren Millionen Personen in über 200 Ländern genutzt wird. Allerdings liegt auch hier kein validiertes Kompetenzmodell zugrunde.

Computerbasierte Diagnoseinstrumente für den Grundbildungsbereich Erwachsener in England sind z. B. der *Target Skills: Initial Assessment*³⁹ und der *Skills for Life Survey*⁴⁰. Beide sind jedoch nicht frei zugänglich und werden daher im Folgenden nur kurz beschrieben.

Target Skills bietet eine computerbasierte Diagnostik und Lernmaterialien in mehreren Dimensionen an: Basic Skills Screener (ein Test bezüglich grundlegender Fertigkeiten in Literalität und mathematischen Grundfertigkeiten), Entry Level sowie Level 1 für Literacy und Numeracy (Lernmaterialien für Erwachsene, die sich auf das Kerncurriculum in den Entry-Level beziehen). Die Materialien sind frei zugänglich, müssen aber käuflich erworben werden. Zum *Target Skills: Initial Assessment* berichteten Brooks u.a. im Report vom NRDC, dass dieser nicht für formative Zwecke geeignet und zudem aufgrund eines möglichen Einflusses der ICT-Literacy nicht ausreichend valide sei (Brooks u. a., 2005, S. 105). Beim *Skills for Life Survey* handelt es sich um eine nationale Studie in England, in der Literalitäts-, Mathe- und Computerkompetenzen erhoben werden (vgl. auch Kap. 2.3.3). Das Instrument ist nicht öffentlich und wird von

³⁸ www.gcflearnfree.org (zuletzt geprüft am 03.11. 2014)

³⁹ www.targetskills.net (zuletzt geprüft am 03.11. 2014)

⁴⁰ <https://www.gov.uk/government/organisations/departement-for-business-innovation-skills> (zuletzt geprüft am 03.11. 2014)

Brooks u. a. hinsichtlich mangelndem *gender* und *cultural mainstreaming*, einem möglichen Einfluss der ICT-Literacy sowie mangelnder Authentizität kritisch bewertet (Brooks u. a., 2005, S. 118). Der Test ist in seinen jeweiligen Abschnitten adaptiv.

Zusammenfassend werden inzwischen Computer in internationalen Vergleichsstudien zur Messung der Literalität (z. B. PIAAC, Skills for Life, IALS) eingesetzt. Frei zugängliche computerbasierte Diagnoseinstrumente liefern jedoch nach dem gegenwärtigen Erkenntnisstand keine Diagnostik, die eine differenzierte und kompetenztheoretisch überprüfte Rückmeldung ermöglichen.

Was insbesondere aus dem Report von Brooks u. a. (2005) hervorgeht, ist der mögliche Einfluss der ICT-Literacy auf das Testergebnis. Um valide Ergebnisse zu erzielen, muss der Einfluss der Computerkompetenz weitergehend ausgeschlossen werden können. Computerbasierte Diagnoseinstrumente sind daher so zu gestalten, dass die Nutzer und Nutzerinnen nicht kognitiv allein aufgrund der Aufgabendarstellung überfordert werden. Hinweise auf mögliche Einflussfaktoren liefern u.a. die Kognitionspsychologie - insbesondere die CLT sowie die CTML (Kap. 2.4) -, die Äquivalenzproblematik (Kap. 2.5) sowie die Usability-Forschung (Kap. 2.6). In den folgenden Kapiteln werden diese möglichen Einflussfaktoren näher erläutert.

2.4 Cognitive Load Theory und Cognitive Theory of Multimedia Learning

Die CLT und die CTML liefern Modelle über die Wirkungsweise von Auslastungsfaktoren auf das Arbeitsgedächtnisses. Aus beiden Theorien lassen sich Überlegungen ableiten, wie computerbasierte Diagnoseinstrumente gestaltet werden können, um den Einfluss der ICT-Literacy auf das Diagnoseergebnis zu minimieren bzw. zu verhindern.

2.4.1 Grundidee und Annahmen der Cognitive Load Theory

Die Grundidee der CLT besteht in der Erforschung der kognitiven Belastung beim Lernen. Sie hat zum Ziel, psychologische Phänomene zu erklären und Hinweise für die Informationsdarstellung zu geben, um die Informationsverarbeitung zu optimieren (Moreno & Park, 2010, S. 9). Lernende sollen darin unterstützt werden, die verfügbaren kognitiven Ressourcen nicht für die Erschließung irrelevanter Informationen des Aufgabenmaterials, sondern für die Verarbeitung von Informationen und das tiefere Verständnis zu nutzen. Wie auch der im Vorfeld bestehenden Theorie des „mental Load“ (Moray, 1979) liegt der CLT die Annahme zugrunde, dass das Arbeitsgedächtnis begrenzt und das Langzeitgedächtnis unbegrenzt sind. In neueren Begriffsbestimmungen repräsentiert der *Mental Load* die Anforderungen an die kognitiven Ressourcen, die aus der Interaktion sowohl zwischen Aufgaben- als auch Subjektcharakteristiken bei einer

Instruktion resultieren (Cook, Zheng & Blaz, 2009, S. 38). In der Theorie von Moray wurden allerdings die psychologischen Effekte auf die Auslastung des Arbeitsgedächtnisses (Einstellungen, Erwartungen, individuelle Ziele) nicht berücksichtigt (Moreno & Park, 2010, S. 10). In der CLT werden hingegen personale bedingt Voraussetzungen einbezogen. Erste Skizzierungen der Theorie und die Verwendung des Begriffs Cognitive Load finden sich in Artikeln von Sweller (Sweller, 1988, 1989), wobei sie schließlich von Chandler und Sweller (1991) konkretisiert wurden. Im Fokus der CLT (CLT) steht das Arbeitsgedächtnis, welches dem Modell von Baddeley (1976) zufolge neben dem Langzeitgedächtnis einen Bestandteil des menschlichen Gedächtnisses darstellt. Das Arbeitsgedächtnis wird benötigt, um aktuelle Prozesse zu bearbeiten, um neues Wissen entweder an bereits Vorhandenes anzuknüpfen oder um neue Schemata zu entwickeln. Es nimmt Informationen über den visuellen/piktographischen oder den auditiven/verbalen Kanal auf. Die Annahmen der CLT werden besonders bei der Gestaltung von Instruktionen im Lernkontext berücksichtigt (Paas, Tuovinen, Tabbers & van Gerven, 2003, S. 63). Ein wesentliches Merkmal ist, dass die Kapazität des Arbeitsgedächtnisses und die Aufnahmefähigkeit begrenzt sind. Als Auslastungsfaktoren des Arbeitsgedächtnisses beim Lernen gelten u. a. das Aufgabenformat, die Aufgabenkomplexität und der Einsatz von Multimedia. Die Auslastung des Arbeitsgedächtnisses ist dabei die subjektiv wahrgenommene Aufgabenschwierigkeit, beeinflusst durch Motivation, Fähigkeit, Erwartung, Training, Zeit, Stress, Ermüdung und Umstände (Kantowitz, 1987, S. 97). Der CLT zufolge lässt sich das Arbeitsgedächtnis in drei Bereiche unterteilen: *Intrinsic Load*, *Extraneous Load* und *Germane Load* (Sweller, 2010, S. 40). Zusätzlich zu den drei Bereichen besteht freie Kapazität (vgl. Abbildung 2).

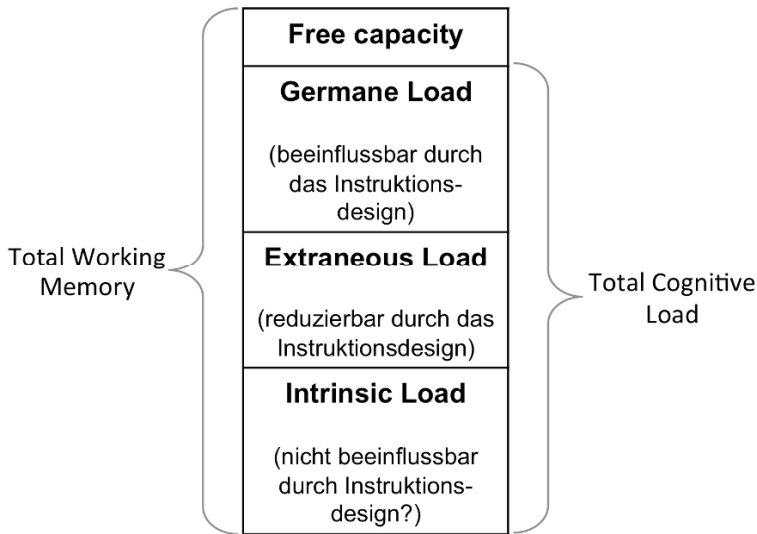


Abbildung 2: Arbeitsgedächtnis und Cognitive Load (Moreno & Park, 2010, S. 18)

Der *Intrinsic Load* ist der Anteil kognitiver Belastung, der in dem intellektuellen Anspruch des Lernmaterials begründet ist. Auslastungsfaktor ist die Element-Aktivität, die durch die Anzahl der gleichzeitig im Arbeitsgedächtnis zu verarbeitenden einzelnen Wissensinhalte entsteht. Lerninhalte, die lediglich den Abruf einzelner isolierter Wissensbestandteile erfordern, wie beispielsweise das Vokabeln lernen, beanspruchen den *Intrinsic Load* nur wenig im Vergleich zu Wissensinhalten, die in einem komplexen Zusammenhang stehen und für das Verständnis des Lerninhalts gleichzeitig im Arbeitsgedächtnis repräsentiert sein müssen (Sweller, 2005). Der *Extraneous Load* entsteht durch den überflüssigen Anteil an Beanspruchung durch eine nicht optimierte Gestaltung des Lernmaterials. Ist das Lernmaterial beispielsweise unübersichtlich gestaltet, müssen kognitive Ressourcen beansprucht werden, um die relevanten Informationen zu identifizieren. Dies würde eine vergleichsweise hohe Auslastung des *Extraneous Load* zur Folge haben. Bei dem *Germane Cognitive Load* handelt es sich um den Anteil kognitiver Beanspruchung, welcher für das Lernen wichtig ist und zu dessen Gunsten es gilt, den *Extraneous Load* möglichst gering zu halten. Der *Germane Load* wird während der tieferen Verarbeitung von Informationen sowie der Bildung neuer oder der Anpassung bereits bestehender Schemata (und somit zum Lernen) beansprucht. Allerdings besteht eine Beanspruchung des *Germane Load* nur, wenn auch der *Intrinsic Load* eine Auslastung erfährt. Ist der *Intrinsic Load*

niedrig und der *Extraneous Load* hoch, ist dennoch Kapazität für den *Germane Load* vorhanden; jedoch kann kein großer Lerneffekt stattfinden, da das Arbeitsgedächtnis vorwiegend mit externen Faktoren (Darstellung der Inhalte) und nicht mit den Inhalten selbst ausgelastet ist. Folglich steigt die Kapazität des *Germane Load*, wenn auch der *Intrinsic Load* steigt (Sweller, 2010, S. 44). Effekte des *Extraneous Load* können daher nur beobachtet werden, wenn auch der *Intrinsic Load* beansprucht wird (Sweller, 2005, S. 28). Der *Intrinsic* und der *Extraneous Load* sind additiv, d. h. wenn die Auslastungsgrenze des Arbeitsgedächtnisses erreicht ist, ist das tiefere Verständnis und somit das Lernen gefährdet. Im Gegensatz zum *Intrinsic Load* sind *Extraneous* und *Germane Load* von dem Aufgabendesign beeinflussbar (Sweller, 2010, S. 44).

Schlussfolgernd hat die Gestaltung der Aufgabe einen direkten Einfluss auf die Kapazitätsauslastung des Arbeitsgedächtnisses und insbesondere auf den *Extraneous Load*. Ist die Auslastung durch das Aufgabenformat hoch, ist weniger Arbeitsspeicher für die Verarbeitung der Informationen vorhanden.

Studien zeigen allerdings, dass die Beanspruchung des Cognitive Load in einem Zusammenhang mit dem Vorwissen, räumlichen Denk- und Vorstellungsvermögen, Selbstregulation sowie Motivation im Allgemeinen stehen (Low & Jin, 2009). Erklären lassen sich die Zusammenhänge zwischen Cognitive Load und dem Vorwissen, dem räumlichen Denk- sowie dem Vorstellungsvermögen mit den Grundprozessen des Lernens: Lernen geschieht, wenn neue Schemata entwickelt werden (Akkommodation) oder neue Elemente in bereits bestehende Schemata integriert werden (Assimilation) (vgl. z. B. Graf, Zimbardo & Gerrig, 2007, S. 66). Die Schemata werden im Langzeitgedächtnis gespeichert und bei Bedarf im Arbeitsgedächtnis präsentiert. Dabei werden sie als einzelne Elemente (*Chunks*) behandelt. Können die Schemata zunehmend automatisiert abgerufen werden, „verschwindet“ zunehmend die Grenze des Arbeitsgedächtnisses, da die Informationen direkt in das Langzeitgedächtnis in bereits vorhandene Schemata integriert werden können (Kirschner, Kester & Corbalan, 2011, S. 1). Das Vorwissen kann wiederum dazu beitragen, die Auslastung des *Extraneous Load* zu reduzieren, indem einzelne Elemente des Informationsmaterials zu größeren Einheiten zusammengefasst werden. Dieser Prozess wird als *chunking* bezeichnet (Miller, 1956). Andersherum kann einer starken Beanspruchung des Arbeitsgedächtnisses entgegengewirkt werden, indem Chunks in kleinere Einheiten gesplittet werden.

Kritisiert wird an der CLT, dass die Motivation, das Vorwissen und auch die Anstrengungsbereitschaft nicht explizit berücksichtigt sind. Auch wenn unbestritten ist, dass die Motivation einen Einfluss auf die Auslastung des Arbeitsgedächtnisses und somit auf den Lernerfolg hat und die Selbstregulationsstrategien und die Motivation das Lernen positiv beeinflussen können (Deci & Ryan,

1993), sind die Erkenntnisse über den Zusammenhang zwischen Motivation und Cognitive Load rar (Low & Jin, 2009, S. 154). Es wäre zu untersuchen, in wie weit eine hohe bzw. niedrige Motivation die Beanspruchung des Arbeitsgedächtnis beeinflussen. Schlussfolgernd gilt es einerseits, die Faktoren zu berücksichtigen und andererseits den Einfluss der Motivation auf die Auslastung des Arbeitsgedächtnisses stärker zu erforschen.

2.4.2 *Cognitive Theory of Multimedia Learning*

Die CTML differenziert den Ansatz der CLT aus und bezieht diesen speziell auf das Lernen mit Multimedia. Der Einsatz von Multimedia ermöglicht die Informationsweitergabe sowohl über den auditiven/verbalen als auch den visuellen/piktographischen Kanal. Einerseits kann diese duale Beanspruchung die Informationsaufnahme positiv beeinflussen bzw. vereinfachen. Andererseits können durch verschiedene Reize das Interesse, die Aufmerksamkeit und die Lernmotivation gesteigert werden (Moreno, 2005, S. 508). Es ist daher davon auszugehen, dass eine Anreicherung von Aufgabenformaten durch Multimedia die Informationsaufnahme (z. B. Instruktion und Fragestellung in einem E-Assessment) erleichtert. Folglich hat dies einen reduzierenden Effekt auf die Auslastung des Cognitive Load, so dass mehr Gedächtniskapazität vorhanden ist, um die Aufgabe zu bearbeiten. Zudem kann die Einbindung von Multimedia die Aufmerksamkeit und Motivation steigern.

Mayers Theorie der kognitiven Belastung beim Lernen mit Multimedia liegen drei Annahmen zu Grunde (Mayer, 2005, S. 33 ff):

1) *Dual Channel Assumption*: Das Arbeitsgedächtnis beinhaltet zwei unterschiedliche Systeme bzw. Kanäle für die Enkodierung und Verarbeitung von visuellen/piktographischen und auditiven/verbalen Inhalten.

2) *Limited Capacity Assumption*: Die Ressourcen für die Repräsentation in den jeweiligen Kanälen (visuell/piktographischen und auditiv/verbal) sind begrenzt. Wenn die Prozesse die Ressourcen überschreiten, entsteht der sog. *Cognitive Overload*. Die Begrenzung der Kanäle führt dazu, dass Entscheidungen darüber getroffen werden müssen, welchen Informationen Aufmerksamkeit geschenkt wird und zu welchem Grad Verknüpfungen zwischen den Informationen bzw. zwischen den Informationen und bereits vorhandenem Wissen hergestellt werden sollen. Hierbei können metakognitive Strategien hilfreich sein.

3) *Active Processing Assumption*: Bedeutungsvolles Lernen findet statt, wenn die Personen aktiv und simultan die Informationen sowohl vom visuellen als auch vom auditiven Kanal verarbeiten. Beim Lernen mit Multimedia laufen dabei fünf Prozesse ab (Mayer, 2005, S. 38): 1. Auswahl relevanter Wörter für die Repräsentation im verbalen Arbeitsgedächtnis, 2. Auswahl relevanter Bilder für die Repräsentation im visuellen Arbeitsgedächtnis, 3. Organisation der aus-

gewählten Wörter in ein verbales Modell, 4. Organisation der ausgewählten Bilder in einem piktographischen Modell und 5. Integration der verbalen und piktographischen Repräsentationen untereinander und mit dem vorhandenen Vorwissen (vgl. z. B. auch J. R. Anderson, 2013).

Für das Designen von Benutzer-/Benutzerinnenoberflächen sind somit visuelle und auditive Informationen aufeinander abzustimmen, so dass die Aufnahme der Informationen weniger Ressourcen beansprucht als wenn die Informationen in ihrer unterschiedlichen Präsentationsform nicht aufeinander abgestimmt und einzeln verarbeitet werden müssten. So können die Informationen schneller und "einfacher" mit bereits bestehenden Modellen abgeglichen und ggf. angepasst (assimiliert) werden.

Dass eine angemessene Darstellung beim multimedialen Lernen die kognitive Belastung reduzieren können, belegen zahlreiche Studien (Mayer & Anderson, 1991; Mayer & Moreno, 2003; Mayer, Moreno, Boire & Vagge, 1999; Mayer & Sims, 1994). Doch führt wiederum eine für den Lernenden nicht adäquate Präsentation multimedialer Inhalte zu einer Steigerung der kognitiven Belastung und behindert somit das Lernen. Zudem darf der Medienreichtum allerdings nicht zu einer „Verkomplizierung“ (*overcomplication*) führen, noch darf es zu einer zu starken „Vereinfachung“ (*oversimplification*) kommen (Schwabe, 2001, S. 55). Gerade in Bezug auf funktionale Analphabeten ist ein geringer Toleranzbereich zu vermuten. Das bedeutet, die Medienwahl muss möglichst zu deren Lese- und Schreibkompetenz sowie zu deren ICT-Literacy passen.

Aus den Annahmen zur CTML sind als Konsequenzen für die Gestaltung von Multimedia insbesondere zwei Prinzipien zu folgern: 1) das präsentierte Material soll eine kohärente Struktur aufweisen; 2) die dargebotene(n) Informationen solle(n) Orientierungsmöglichkeiten geben, um sich als Rezipient die Struktur erschließen zu können. Diese Prinzipien finden sich auch in Gestaltungsrichtlinien für Usability wieder (vgl. Kap. 2.6.3.2, z. B. „Gesetz der Nähe“).

Kritisiert wird an der CTML, dass die Motivation des Lernens außer Acht gelassen wird, welche eine einflussreiche Rolle spielt (Merriënboer & Sweller, 2005). Eine Reduktion der kognitiven Belastung führt nur zu einem Lerneffekt, wenn die lernende Person motiviert ist zu lernen, die freien Ressourcen zu nutzen und in den *Germane Load* zu investieren. Problematisch an diesem Konzept der Kognitiven Belastung ist zudem, dass von einer kumulativen und statischen Beschaffenheit ausgegangen wird. Beim Lernen – und insbesondere beim komplexen Lernen – sind mehrere Vorgänge und Entscheidungen involviert, welche die darauf folgenden lernrelevanten Aspekte beeinflussen. Und so erwähnen Cook u. a., dass auch der *Cognitive Load* als dynamisches Konstrukt betrachtet werden müsse (Cook u. a., 2009, S. 38). Lösungsansätze liefern beispielsweise

Xie & Salvendy (2000) mit der Differenzierung zwischen *instantaneous Load*, *peak Load*, *accumulated Load*, *average Load* und *overall Load*.

Die CLT und die CTML unterscheiden sich darin, dass bei der CLT die Interaktionen von Informationen und weniger der Einsatz von Multimedia im Vordergrund stehen, wohingegen bei der CTML der Fokus auf mentalen Repräsentationen durch Multimedia liegt. Mayer sieht den Unterschied zudem darin, dass bei der CLT die 5 Prozesse (Auswahl relevanter Wörter/Bilder für die Repräsentation im verbalen/visuellen Arbeitsgedächtnis, Organisation der ausgewählten Wörter/Bilder in ein verbales/piktographischen Modell, Integration der verbalen und piktographischen Repräsentationen untereinander und mit dem vorhandenen Vorwissen) nicht berücksichtigt werden (Brünken, Plass & Leutner, 2004).

Bisher haben die Ansätze der CLT und CLML nur im Kontext von Lernen und Lehren mit Multimedia stattgefunden, nicht aber im Bereich der Kompetenzdiagnostik im Allgemeinen und des E-Assessments in Verbindung mit der probabilistischen Testtheorie (IRT – Item Response Theory)⁴¹ im Besonderen.

Es ist davon auszugehen, dass eine hohe Usability (vgl. Kap. 2.6) eine geringe Auslastung des *Extraneous Load* zur Folge hat. Anders herum formuliert: Eine schlechte Usability würde den Nutzer/die Nutzerin kognitiv stark beanspruchen – beispielsweise kann die Orientierung auf einer Website schwierig sein, wenn diese unübersichtlich gestaltet ist (vgl. Tabelle 5).

⁴¹ Die Aufgaben der Online-Testumgebung wurden auf der Basis der IRT entwickelt. Eine genauere Erläuterung hierzu erfolgt in Kap. 4, in dem die Entwicklung der Online-Testumgebung und der Projektkontext erläutert werden.

Usability Auslastung	ausgeprägte Usability ohne Einsatz von Multimedia	ausgeprägte Usability mit Einsatz von Multimedia	Usability niedrig (mit/ohne Einsatz von Multimedia)
Extraneous Load	mittel	niedrig	hoch
Intrinsic Load	hoch	hoch	hoch
Germane Load	hoch	niedrig	niedrig

Tabelle 5: Angenommener Zusammenhang zwischen der Beanspruchung des Cognitive Load und Usability

Die kognitive Beanspruchung würde vermutlich umso stärker ausfallen, je weniger Computererfahrung die Personen haben. Personen mit wenig Computererfahrung haben bisher wenig mentale Modelle bzw. Schemata hinsichtlich Benutzer-/Benutzerinnenoberflächen (z. B. hinsichtlich eines typischen Seitendesigns und/oder Seitenstruktur sowie grundlegender Funktionen) gebildet, die sie abrufen könnten. Daher ist der *Extraneous Load* bei Personen mit wenig Computererfahrung und -kompetenz vermutlich stärker ausgelastet als bei Personen mit mehr Computererfahrung und -kompetenz (Koppel, 2011; Mayer, 2005; Plass, Moreno & Brünken, 2010). Der *Extraneous Load* würde aufgrund der Computerkompetenz und -erfahrung sowie möglicherweise auch aufgrund der Seitengestaltung zu Lasten des *Germane Load* stark ausgelastet sein und durch eine gleichzeitige Repräsentation von komplexen Wissensbestandteilen die Lösung der Aufgabe beeinträchtigen.

2.4.3 Messmöglichkeiten des Cognitive Load

Wie bereits oben beschrieben ist eine zentrale Annahme, dass das Instruktionsdesign und oder die Methoden entweder eine das Lernen unterstützende oder eine dem Lernen hinderliche (durch eine hohe Beanspruchung des *Extraneous Load*) Auslastung zur Folge hat. Eine weitere Annahme besagt, dass der Cognitive Load in Abhängigkeit der Aufgabenkomplexität variiert und davon auch die Beanspruchung des *Intrinsic Load* abhängt (Brünken, Seufert & Paas, 2010, S. 181). Die bisherigen Erläuterungen und Annahmen deuten bereits darauf hin, dass der Cognitive Load nicht direkt beobachtbar ist, sondern nur indirekt gemessen werden kann. In den vergangenen 15 Jahren konnten weder standardisierte Messmethoden noch ein „allgemeingültiges“ Forschungsparadigma entwickelt werden. Die bisherigen Messmethoden werden in analytische, aufgaben- und performanzbasierte Methoden sowie physiologische Techniken und Beurteilungsskalen kategorisiert. In Anlehnung an Brünken u. a. (2010) können diese wiederum in zwei Bereiche eingeteilt werden:

1) Subjektive Methode bzw. Selbsteinschätzung: Personen werden nach ihrer wahrgenommenen Auslastung befragt;

2) Objektive Messmethoden: Es wird die physiologische Performanz gemessen.

Zu 1) Die subjektive Messmethode werden Personen gebeten, ihre wahrgenommene Gedächtnisauslastung auf einer Skala zwischen „gar nicht“ und „sehr viel“ einzuschätzen. Die Skalenmethode basiert auf der Annahme, dass die Personen eine valide Einschätzung bezüglich ihrer Gedächtnisleistung und der zu bewältigenden Aufgabe abgeben können. Die meisten Methoden nutzen eine 7-9-stufige Likert-Skala. Kombiniert werden die Methoden oftmals mit der Einschätzung der Aufgabenschwierigkeit. Bisherige Studien weisen eine hohe Korrelation zwischen den beiden Dimensionen (Selbsteinschätzung der Gedächtnisauslastung sowie wahrgenommene Aufgabenschwierigkeit) auf, so dass auch eindimensionale Untersuchungen durchaus valide Ergebnisse produzieren. Der größte Vorteil dieser Methoden liegt in dem geringen Aufwand der Umsetzung. Allerdings weist diese Methode auch bedeutsame Grenzen auf, denn Personen schätzen im *Nachhinein* die wahrgenommene Gedächtnisleistung ein (Brünken u. a., 2010, S. 183). Es ist nicht ersichtlich, ob bzw. welche Faktoren diese Selbsteinschätzung beeinflussen. Diese Faktoren können sich sowohl durch Rahmenbedingungen (z. B. Lautstärke in dem Untersuchungsraum) als auch durch physiologische Gegebenheiten (z. B. Müdigkeit) generieren. Andererseits kann dieser Effekt ausgeschlossen werden, wenn zu unterschiedlichen Zeitpunkten eines Tests die Selbsteinschätzung wiederholt werden. Eine weitere größere Einschränkung ergibt sich aus der Inhaltsvalidität: Für die Personen, die eine Selbsteinschätzung vornehmen ist meist nicht deutlich zu differenzieren, welcher Bereich des Cognitive Load ausgelastet ist (Brünken u. a., 2010, S. 183). Ein Ansatz für die Messung des *Extraneous Load* kann sein, Aufgaben mit gleicher Aufgabenschwierigkeit aber unterschiedlich dargebotenen Formaten bearbeiten zu lassen.

Zu 2) Es existieren zahlreiche Indikatoren zur objektiven Messung des *Cognitive Load*. Bezüglich des Lernprozesses können diese in „Outcome-Variablen“, „Input-Variablen“ sowie „Prozessorientierte Variablen“ (z. B. Performanz oder Bearbeitungszeit) differenziert werden, wobei der objektivste Indikator der Lernzuwachs ist (Brünken u. a., 2010, S. 183). Messmöglichkeiten bestehen in experimentell kontrollierten Lernsituationen. Fällt der Lernzuwachs unterschiedlich aus, können diese Unterschiede in der unterschiedlichen Auslastung des *Cognitive Load* begründet sein. Grenzen dieser Methode bestehen darin, dass zwar die abhängige Variable Lernzuwachs und die unabhängigen Variablen der experimentellen Bedingungen benannt werden können. Aber es kann nicht eindeutig festgestellt werden, ob tatsächlich der *Cognitive Load* den Effekt auf

den Lernzuwachs auslöst oder ob andere Einflussfaktoren eine Rolle spielen und die Ursachen für diese Varianz sind (z. B. die Motivation). Der Lernzuwachs selbst stellt somit keinen erfüllenden und zureichend erklärenden Indikator dar, dennoch kann der Lernzuwachs in Kombination mit weiteren Messmethoden hinreichende Validität liefern.

Cook u. a. (2009, S. 42) stellen die Messmethoden in einer Matrix dar und bewerten sie hinsichtlich ihrer Subjektivität/Objektivität und der Auslastung, die durch die Methode gemessen werden kann. Dabei ordnen sie ebenfalls die Beurteilungsskalen und analytischen Vorgehen den subjektiven und Performanz sowie physiologische Messungen (beispielsweise der Herzfrequenz oder Gedächtnisaktivität) den objektiven Messmethoden zu. Allerdings legen sie nicht die Differenzierung zwischen *Intrinsic* und *Extraneous Load*, sondern die von Xie & Salvendy (2000) zugrunde (*instantaneous Load*, *peak Load*, *accumulated Load*, *average Load* und *overall Load*, vgl. vorheriges Kap.). Die Wahl der Messmethoden hängt sowohl vom Kontext als auch von der Fragestellung ab. Kontextbedingt ist der Einsatz bestimmter Verfahren aufgrund der Zielgruppe oder der zur Verfügung stehenden Ressourcen ausgeschlossen. Beispielsweise bedeutet der Einsatz von Instrumenten für die Messung der Gedächtnisaktivität einen hohen finanziellen Aufwand und ist für Forschende nicht immer tragbar. Schließlich sei angemerkt, dass es sich bei dem Begriff Cognitive Load um ein Konstrukt handelt, welches immer nur indirekt gemessen werden kann. Eine direkte Beobachtung ist nicht möglich. Dennoch liefern die CLT und CTML wesentliche Hinweise für die Gestaltung von Benutzer-/Benutzerinnenoberflächen und somit auch für die Gestaltung von computerbasierten Instrumenten. Insbesondere die Annahmen über den *Intrinsic* und den *Extraneous Load* weisen darauf hin, dass die Gestaltung der Benutzer-/Benutzerinnenoberflächen von computerbasierten Diagnoseinstrumenten einen Einfluss auf die Performanz und daher auch auf das zu erfassende Konstrukt ausüben kann. Insbesondere bei einer direkten Übertragung von einer papier- in eine computerbasierte Version muss kritisch hinterfragt werden, ob sich durch die Darstellungsform der Items das Testergebnis verändert, worauf im folgenden Kapitel eingegangen wird.

2.5 Exkurs: Äquivalenzproblematik

Die Äquivalenzproblematik behandelt die Schwierigkeit eines Transfers von einer papier- auf eine computerbasierte Testung.

„There is one methodological issue that should be considered from a technological point of view, however, and this is validity. Different validity issues

may arise when TBA [Technology Based Assessment] is applied to replace traditional paper-based assessment and when skills related to the digital world assessed“ (Csapó u. a., 2012, S. 145).

Zu Beginn der Einsatzphase von computerbasierter Diagnostik wurden oftmals bereits bestehende papierbasierte Instrumente in eine technologiebasierte Version transferiert. Doch wenn der Testmodus geändert wird – z. B. von papierbasiert (PP für *Paper Pencil*) zu computerbasiert – kann sich unter Umständen die Natur des erfassten Konstrukts verändern, so dass zwischen den beiden Versionen keine Äquivalenz mehr besteht.

Eine Studie von Pomplun, Ritchie & Custer (2006) deutet auf die Äquivalenzproblematik hin: Trotz eines 1:1-Transfers von einem papier- zu einem computerbasierten Erhebungsinstrument wurden unterschiedliche Effekte erzeugt. In der Untersuchung wurden Faktoren extrahiert und untersucht, welche auf unterschiedliche Effekte einer technologiebasierten respektive papierbasierten Befragung hinweisen. Relevante Faktoren beziehen sich auf Probanden-/ Probandinnen- und Itemcharakteristikvariablen. Die Studie wurde bei Grundschüler und -schülerinnen durchgeführt, welche zu dem Zeitpunkt der Testdurchführung niedrige Lese- und Schreibfähigkeiten hatten. Die Ergebnisse hinsichtlich des Itemformats zeigen, dass Probanden und Probandinnen in der papierbasierten Befragung einen höheren Wert erzielten, als im computerbasierten Assessment (Pomplun u. a., 2006, S. 135). Zudem fielen die Effekte um so größer aus, je weniger die Person mit dem Computer gearbeitet hat (Pomplun u. a., 2006, S. 129). Zusammenfassend zeigt diese Studie, dass Differenzen im Antwortverhalten aus den Unterschieden der Erhebungsformen (PP vs. computerbasiert) resultieren und in der ICT-Literacy begründet sein können.

Auch die Studie von Bennett (2008) belegt, dass Testergebnisse im computerbasierten Assessment niedriger als im Vergleich zu den Ergebnissen der papierbasierten Erhebung ausfielen und das Maß der Vertrautheit im Umgang mit Computern eine Rolle spielt. Weiterhin wurde deutlich, dass der Unterschied zwischen PP-Erhebungen und computerbasierten Assessments in Abhängigkeit zum Itemformat steht: Der Effekt auf das Testergebnis war größer, wenn Befragungsteilnehmende Sätze konstruieren mussten, als der Effekt von Multiple-Choice Items im computerbasiert-PP-Vergleich (Bennett u. a., 2008, S. 25 f.).

Deutlich wird einerseits, dass papierbasierte und computerbasierte Befragungen trotz gleicher Items nicht immer automatisch äquivalent sind. Andererseits zeigt die CTML, dass der Einsatz von Multimedia sehr zur Erklärung und zum Verständnis verschiedener Inhalte beitragen kann. Da durch die technische Entwicklung sich inzwischen die Formate in der computerbasierten Diagnostik

stark von denen im PP eingesetzten Itemformaten sowie früheren Formaten computerbasierter Diagnostik unterscheiden (können), ist entweder eine genaue Prüfung der Äquivalenz⁴² notwendig oder es ist von Beginn der Testkonstruktion zu entscheiden und berücksichtigen, in welchem Testmodus die Daten erhoben werden sollen. Des Weiteren sind Charakteristiken der potenziellen Zielgruppe einzubeziehen: Haben potenzielle Nutz und Nutzerinnen Erfahrung im Umgang mit Computern? Arbeiten sie gerne mit dem Computer? Stehen sie dem Einsatz von Computern in Testkontexten kritisch gegenüber? Der mögliche Einfluss dieser Faktoren (insbesondere der ICT-Literacy) auf die Performanz ist so weit wie möglich zu kontrollieren. Im Vorfeld ist daher bei jeglicher Gestaltung von computerbasierter Diagnostik auf eine zielgruppenadäquate Usability zu achten. Zusammenfassend fokussiert die Äquivalenzproblematik die Inhaltsvalidität von Datenerhebungsinstrumenten in unterschiedlichen Erhebungsformen. Im Rahmen dieses Forschungsvorhabens wird die Frage der Äquivalenz zwischen der papier- und der computerbasierten Diagnostik nicht weiter behandelt, da die Entwicklung und Evaluation hinsichtlich der Usability im Vordergrund stehen. Es wird zwar davon ausgegangen, dass durch eine ausgeprägte Usability auch die Inhaltsvalidität steigt. Doch ist die Überprüfung der Inhaltsvalidität erst nach Abschluss der Entwicklung möglich bzw. sinnvoll.

2.6 Usability

Bei der Gestaltung von Webseiten im Allgemeinen und computerbasierten Diagnoseinstrumenten im Besonderen sind neben zielgruppenspezifischen Voraussetzung und kognitionspsychologischen Aspekten. Die Anforderungen an eine ausgeprägte Usability werden über die Formulierung von Richtlinien konkretisiert. Bevor verschiedene Richtlinien für die Usability vorgestellt werden, wird der Begriff Usability mit seinen verschiedenen Bedeutungszuschreibungen vorgestellt, um schließlich eine Definition sowie Abgrenzung zu verwandten Begriffen vorzunehmen.

2.6.1 Usability – eine Begriffsan- und -einordnung

„Usability is a quality attribute that assesses how easy user interfaces are to use. The word "usability"

⁴² Guidelines der International Testing Commission (ITC) konstatieren, dass Äquivalenz zwischen computerbasierten und papierbasierten Tests dann besteht, wenn beide Tests vergleichbar reliabel sind, ausreichend hoch miteinander korrelieren, in ähnlichem Ausmaß mit anderen Testverfahren und externen Kriterien korrelieren sowie vergleichbare Mittelwerte und Standardabweichungen aufweisen (Jurecka & Hartig, 2007, S. 42).

also refers to methods for improving ease-of-use during the design process⁴³“ (Nielsen, 2012).

Der Begriff Usability wird im Deutschen meist mit Gebrauchstauglichkeit und/oder Benutzer-/Benutzerinnenfreundlichkeit übersetzt. Synonyme sind Benutzer-/Benutzerinnenfreundlichkeit, Bedienbarkeit und Brauchbarkeit. Allein die Existenz mehrerer Synonyme deutet auf eine nicht trennscharfe Verwendung des Begriffs im deutschen Sprachraum hin. Mit dem Begriff Gebrauchstauglichkeit wird stärker die Funktion und somit das Produkt, mit dem Begriff Benutzer-/Benutzerinnenfreundlichkeit stärker der Nutzer/die Nutzerin und somit eine Serviceorientierung fokussiert (Rampl, 2007). Da im Deutschen kein eindeutiges Äquivalent für den englischen Begriff Usability existiert, weil der englische Begriff Usability auch im Deutschen hinreichend geläufig ist und weil der Begriff Usability hinreichend definiert ist, wird im Folgenden der Begriff englischen Ursprungs *Usability* verwendet. Es existieren mehrere Definitionen des Begriffs Usability, die durch unterschiedliche Fokussierungen auch unterschiedliche Bedeutungsprägungen mitführen.

„Only by defining the abstract concept of „usability“ in terms of these more precise and measurable components can we arrive at an engineering discipline where usability is not just argued but is systematically approached, improved, and evaluated (possibly measured). [...] Clarifying the measurable aspects of usability is much better than aiming at a warm, fuzzy feeling of „user friendliness“ (Shackel, 1991, S. 24).

Um die Usability eines Produktes evaluieren und Usability-Anforderungen für eine bestimmte Zielgruppe entwickeln zu können, ist es unumgänglich, eine Definition herauszuarbeiten, die im Kontext dieser Arbeit Gültigkeit besitzt. Im Folgenden wird das Konstrukt Usability erörtert, um intersubjektiv nachvollziehbar dazustellen und zu begründen, welche Definition schließlich in der vorliegenden Arbeit Verwendung findet.

Der Begriff Usability wurde Mitte der 1980er Jahre eingeführt, um die Bezeichnung *user friendly* (engl. für nutzer-/nutzerinnenfreundlich) zu beschreiben. Der Anspruch einer möglichst erschöpfenden Bedeutung des Begriffs erfordert die Annäherung aus mehreren Richtungen. Wird der Wortstamm betrachtet,

⁴³ Das Zitat stammt von seiner Website <http://www.nngroup.com/articles/usability-101-introduction-to-usability/> (zuletzt geprüft am 03.11. 2014).

erhält er eine erste Schärfung: Der englische Begriff setzt sich aus den Begriffen *use* (engl. für *nutzen*) und *ability* (engl. für *Fähigkeit*, *Geschicklichkeit*, *Leistungsfähigkeit*) zusammen. Demnach bedeutet der Begriff, die *Fähigkeit etwas zu nutzen*. Diese Deutung impliziert, dass Webseiten oder Software ohne Usability sinnlos, weil nicht brauchbar, wären. Nach Rampl (2007) reicht allerdings das reine Erkennen der Notwendigkeit zu kompromisslosen Ausrichtung an den Bedürfnissen der Nutzer und Nutzerinnen nicht aus, sondern der Begriff impliziert ebenso eine Handlungsaufforderung: „*use the ability*“ (engl. für „*Nutze die Möglichkeiten*“). Durch den Imperativ ist somit nicht nur eine Serviceorientierung, wie der Begriff Benutzer-/Benutzerinnenfreundlichkeit vermuten lässt, sondern auch eine Notwendigkeit zu erkennen. Dies lenkt den Fokus etwas weg von der nutzer-/nutzerinnenzentrierten Perspektive und stärker hin zu einer funktionsorientierten Sichtweise.

Wird nun auch berücksichtigt, dass Normen und Standards für die Usability festgelegt sind, ist zudem eine weniger subjektiv abhängige und stärker objektiv feststellbare Ausprägung erkennbar. Somit prägen sowohl die Erfahrungen der Nutzer und Nutzerinnen als auch die Funktion der Technik die Ausprägung der Usability. An dieser Stelle lässt sich bereits erkennen, dass Usability ein relationaler Begriff ist: Die Usability ist immer vor dem Hintergrund des Gebrauchsziels der nutzenden Person zu betrachten: Ein Produkt dient dazu, dass ein Nutzer/eine Nutzerin eine bestimmte Aufgabe erfüllen möchte. Dabei kann es sich beispielsweise um eine Arbeitsaufgabe, ein Spiel oder einen Kaufprozess handeln. Schließlich laufen diese Erläuterungen auf vier Faktoren hinaus, die den Qualitätsgrad der Usability von Benutzer-/Benutzerinnenschnittstellen beeinflussen: a) Der Nutzer/Die Nutzerin, b) die Aufgabe, c) das System und d) die Umgebung bzw. der Anwendungskontext. So lautet eine frühe Definition von Usability:

„[Usability is] the ease of use and acceptability of a system or product for a particular class of users carrying out specific tasks in a specific environment; where ‘ease of use’ affects user performance and satisfaction, and ‘acceptability’ affects whether or not the product is used“ (Bevan, Kirakowski & Maissel, 1991, S. 2).

Usability ist die Einfachheit der Nutzung und die Akzeptanz von einem Produkt, welche Auswirkungen auf die Effizienz und Zufriedenheit einer Produktnutzung haben; die Akzeptanz hat wiederum Einfluss darauf, ob das Produkt genutzt wird. Diese Definition macht insbesondere die Relationalität des Konstrukts Usability deutlich.

Für eine weitere Priorisierung der Faktoren innerhalb des Konstrukts kann die Definition von Nielsen herangezogen werden. Nielsen (1993) geht von der *Usefulness* (engl. für *Brauchbarkeit*) aus, welche beschreibt, ob ein System genutzt werden kann, um ein bestimmtes Ziel zu erreichen. Die *Usefulness* lässt sich wiederum in *Utility* und *Usability* unterteilen. *Utility* bezeichnet dabei, ob ein System für die Anliegen genutzt werden kann, für die es gedacht ist. Die *Usability* ist wiederum der Grad an Qualität, in welcher der Nutzer/die Nutzerin die Interaktion mit etwas erlebt (Nielsen, 1993, S. 25). In diesem Kontext sind die Attribute für die Qualität Erlernbarkeit, Effizienz, Erinnerbarkeit, Fehlertoleranz und Zufriedenheit (vgl. zur näheren Beschreibung der Attribute Kap. 2.6.4). Diese Bezeichnung bezieht – im Gegensatz zu der deutschen Übersetzung – das Erleben des Nutzers/der Nutzerin stärker ein. *Usability* ist somit keine eindimensionale Eigenschaft eines Nutzer-/Nutzerinneninterfaces.

Die *International Organization for Standardization* (ISO) und das Deutsche Institut für Normung (DIN) geben für *Usability* folgende Definition vor⁴⁴:

„[*Usability* ist] das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und mit Zufriedenheit zu erreichen“ (DIN EN ISO, 2011, S. 38).

Diese Definition geht über die von Nielsen (s.o.) hinaus. Nielsen legt den Fokus auf die Leichtigkeit und Unkompliziertheit der Produktnutzung. Der DIN Norm ist diese Anforderung implizit, sie liefert darüber hinaus ein differenziertes Bild, indem Kontextabhängigkeit (Nutzungskontext), die Nutzbarkeit der Technologie (effektiv und effizient) und die Emotionalität des Nutzers/der Nutzerin berücksichtigt werden. Die *Usability* bezieht sich auf die Zeitspanne während der Nutzung eines Produktes. Das bedeutet nicht, dass die *Usability* während der Entwicklung – also vor der Nutzung – keine Rolle spielt, sondern dass die *Usability* nur während der Nutzung beobachtbar und messbar ist.

Da die Definition der ISO bzw. DIN einerseits hinreichend differenziert und andererseits auch eine der geläufigsten ist, wird dem Begriff *Usability* die Definition der ISO bzw. des DIN im weiteren Verlauf der Arbeit zu Grunde gelegt. Ein weiterer Begriff, der fälschlicherweise in manchen Fällen mit *Usability* gleichgesetzt wird, ist der Begriff *Software-Ergonomie*. Die *Software-Ergonomie* ist ein der *Usability* übergeordneter Begriff und schließt diese ein. Die *Software-Ergonomie* besteht aus den drei Bereichen *Gestaltungsgrundsätze*

⁴⁴ Die Definition der ISO und des DIN werden in Kap. 2.6.4 wieder aufgegriffen und eingehend erläutert.

und Rahmenbedingungen, Multimedia-Navigation und Steuerung sowie Auswahl und Kombination relevanter Medien (die Anforderungen an die Software-Ergonomie wurden vom Institut für Normung in der Richtlinie ISO 14915 formuliert). Hier gerät wiederum die Perspektive des Nutzers/der Nutzerin an den Rand des Blickfeldes.

Die Erläuterung des Begriffs Usability sowie die für die Usability relevanten verwandten Konzepte macht deutlich, dass der Begriff Usability einerseits nicht isoliert betrachtet werden kann. Andererseits ergeben sich sowohl Überschneidungen als auch hierarchische Beziehungen zwischen den Konzepten. Die Abbildung 3 dient dazu, das Beziehungsgeflecht der Begrifflichkeiten zu verdeutlichen. Die Abbildung 3 kann nicht den Anspruch der Vollständigkeit erfüllen, wohl aber als Gerüst für den weiteren Verlauf dieser Arbeit dienen.

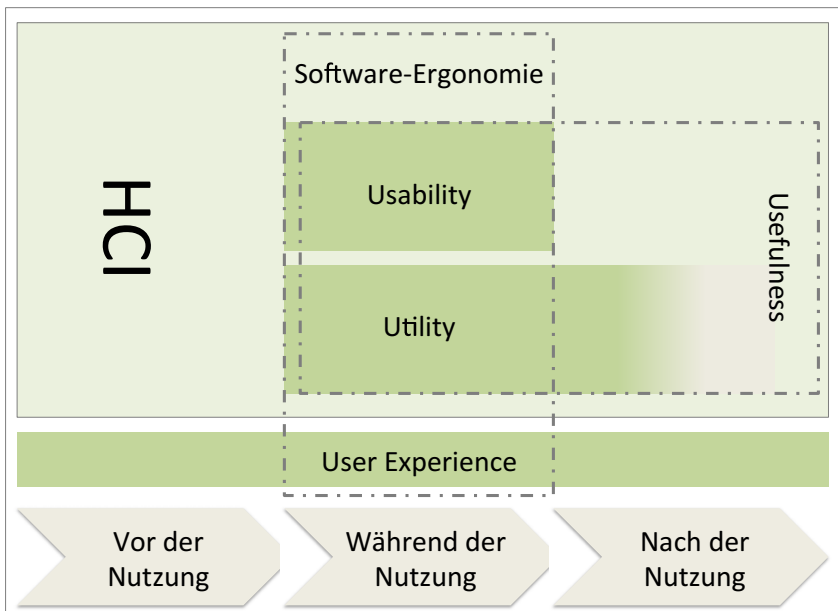


Abbildung 3: Usability in Beziehung zu angrenzenden/verwandten Begriffen

In dieser Abbildung sind die Begriffe unter Berücksichtigung des Zeithorizonts in Beziehung gesetzt. Die Usability ist ein Konstrukt der Human-Computer Interaction (HCI) (Mensch-Computer-Interaktion), in der es um benutzer-/benutzerinnenfreundliches Design von Computersoftware oder Webseiten geht (Preece, 1994). Die HCI bezieht sich auf den Zeitraum vor, während und nach

der Nutzung eines Produktes. Ein Bestandteil der HCI ist die Software-Ergonomie, welche die Begriffe *Usability* und *Utility* einschließt. *Usability* und *Utility* ergeben zusammen die *Usefulness* eines Produktes. Die *Utility* und *Usefulness* gehen über den gegenwärtigen Zeitpunkt der Nutzung hinaus, indem auch der längerfristige – zukünftige Nutzen – mitgedacht wird.

Ein Aspekt, der ebenfalls über den gegenwärtigen Zeitpunkt hinaus geht und zusätzliche auch die Vergangenheit einschließt, ist die *User Experience* (UE). Diese gibt Aufschluss über die Wahrnehmung eines Produkts aus der Perspektive der Nutzer und Nutzerinnen und wird oftmals in Zusammenhang mit der *Usability* überprüft. Mit Hilfe von Erkenntnissen über die UE kann die *Usability* eines Produktes gesteigert werden. Kenntnisse über die UE sind insbesondere hilfreich, wenn über die potenzielle Nutzer-/Nutzerinnengruppe hinsichtlich der Computernutzung und -erfahrung wenig bekannt ist. Da dies bezüglich der Forschungsfrage der Fall ist, soll die UE einer genaueren Betrachtung unterzogen und mit der *Usability* in Beziehung gesetzt werden.

2.6.2 *User Experience (UE)*

Ins Deutsche übersetzt bedeutet UE zunächst Nutzer-/Nutzerinnenerfahrung. Sie wird definiert als „Wahrnehmungen und Reaktionen einer Person, die aus der tatsächlichen und/oder der erwarteten Benutzung eines Produktes, eines Systems oder einer Dienstleistung resultieren" (DIN EN ISO, 2009) und ist auf *interaktive* Computersysteme bezogen (Herczeg, 2009, S. 7).

Damit liegt der Fokus auf den Erfahrungen der Nutzer und Nutzerinnen hinsichtlich eines Produkts. Die Erfahrungen sind allerdings nicht nur abhängig von dem gegenwärtig genutzten System und der zu bearbeitenden Aufgabe, sondern auch von den Erfahrungen, welche der Nutzer/die Nutzerin mitbringt (z. B. haben Nutzer und Nutzerinnen, die viele positive Erfahrungen mit Computern gemacht haben sicherlich ein anderes Erleben während der Nutzung eines Produktes als Personen, die bisher wenig und negative Erfahrungen gemacht haben). Sie ist insbesondere zu berücksichtigen, wenn ausschließlich der Nutzer/die Nutzerin über den Verbleib auf oder das Verlassen einer Website entscheidet, also freiwillig ein Produkt nutzt (Rampl, 2007).

Die UE bekam durch den vermehrten Einsatz in unterschiedlichsten Lebenssituationen eine zunehmende Bedeutung. Die Nutzer und Nutzerinnen erwarteten mehr als ein funktionierendes System, nämlich Unterhaltung. Die UE ist einer der neuesten Begriffe in dem Kriterienset, nach welchem ein System evaluiert werden soll (Petrie & Bevan, 2009, S. 5). Erstmals findet sich der Begriff UE 1998 in einer Veröffentlichung von Norman, der die Berücksichtigung von Aspekten verlangte, die über Effektivität, Effizienz und Zufriedenheit hinausgehen (Norman, 1998).

Die UE umfasst demnach alle Effekte auf den Nutzer/die Nutzerin vor und nach der Nutzung des Produktes. Die Usability hingegen bezieht sich nur auf die Zeitspanne während der Nutzung. Die vorherigen Erfahrungen und die Emotionalität des Nutzers/der Nutzerin werden – wie bereits oben erläutert – weniger berücksichtigt. So kann beispielsweise die UE hoch, die Usability aber niedrig sein. Aufgrund der Einflussfaktoren Mensch, Technologie, Aufgabe und Anwendungskontext ist auch die UE – wie Usability – ein relationaler Begriff.

Schlussfolgernd hängt von den Fähigkeiten und der Wahrnehmung der Nutzer und Nutzerinnen, der Technologie, der zu bearbeitenden Aufgabe und dem Anwendungskontext ab, welche Erfahrungen die Nutzer und Nutzerinnen mit einem Produkt machen und ob ein Produkt als gebrauchstauglich und nutzer-/nutzerinnenfreundlich bewertet wird. Da in dieser Arbeit der Fokus auf den Nutzern und Nutzerinnen liegt und der Frage nachgegangen wird, welche zielgruppenspezifischen Anforderungen hinsichtlich der Usability ein Produkt erfüllen sollte, wird im Folgenden näher erläutert, welche menschlichen Faktoren bei der gebrauchstauglichen und benutzer-/benutzerinnenfreundlichen Gestaltung zu berücksichtigen sind.

2.6.3 Kognitive und physiologische Voraussetzungen

Die kognitiven und physiologischen Voraussetzungen bestimmen, wie Menschen Informationen individuell wahrnehmen und verarbeiten. Die Informationen werden vorerst physisch wahrgenommen und ausgehend von den individuellen mentalen Modellen kognitiv verarbeitet. Die kognitiven und physiologischen Voraussetzungen stellen somit eine zu beachtende Einflussgröße im Kontext der Usability dar.

Im Kap. 2.4 wurde mit der CLT bereits erläutert, dass die Auslastung des Arbeitsgedächtnisses sowohl von der Darbietungsform als auch vom Inhalt abhängig ist. Ob Informationen das Arbeitsgedächtnis stark oder weniger stark auslasten hängt davon ab, wie anstrengend und schwierig es für eine Person ist, sich den Inhalt zu erschließen. Grundlegend wird dies von den physiologischen und kognitiven Voraussetzungen bestimmt. Zu den kognitiven Prozessen zählen u. a. die Wahrnehmung, die Aufmerksamkeit, das Gedächtnis sowie Wissensrepräsentationen und mentale Modelle⁴⁵. Bei der Nutzung von Software bestehen nach Norman (1996) zwei wesentliche Hürden: Der *gulf of execution* (Ausführungskluft) und der *gulf of evaluation* (Auswertungskluft) (Norman, 1996). Der *gulf of execution* beschreibt die Situation, dass der Nutzer/die Nutzerin dem System seine Intention vermitteln muss – beispielsweise mit Hilfe der Eingabe

⁴⁵ Weitere kognitive Prozesse – beispielsweise höhere Hirnleistungen wie Rechnen und Erinnern – werden nicht weiter thematisiert, da diese nicht zwangsläufig eine Rolle bezüglich der Usability spielen.

über die Tastatur oder das Anklicken mit der Computermaus. Bei dem *gulf of evaluation* geht es darum, dass die Ausgaben des Systems von der nutzenden Person korrekt interpretieren und in ihre Handlungen integrieren muss. Für Beides sind sowohl kognitive als auch physiologische Voraussetzungen zu berücksichtigen.

Im Folgenden werden die kognitiven und physiologischen Voraussetzungen sowie deren Rolle bezüglich der Usability kurz erläutert. Bei der Darstellung der kognitiven und physiologischen Grundlagen wird von den Voraussetzungen bei Personen ausgegangen, die weder kognitiv noch physisch eingeschränkt sind.

2.6.3.1 Aufmerksamkeit

Das Thema Aufmerksamkeit wird kontrovers diskutiert (Sarodnick & Brau, 2011, S. 57) und soll hier nur so weit erläutert werden, wie es zum weiteren Verständnis dient.

Aufmerksamkeit ist die auf die Betrachtung eines Objekts gerichtete Bewusstseinshaltung, durch die das Beobachtungsobjekt bewusst wahrgenommen wird (Bergius, 2004, S. 84f). Die Aufmerksamkeit ist begrenzt und legt fest, welche Ereignisse uns bewusst werden. Beispielsweise kann zwischen visueller, auditiver und zentraler Aufmerksamkeit unterschieden werden (J. R. Anderson, 2013), was zu Folge hat, dass sie situationsabhängig auf einzelne Reize (selektive Aufmerksamkeit) oder auf mehrere Reize gleichzeitig gelenkt (geteilte Aufmerksamkeit) wird.

Bezüglich der visuellen Aufmerksamkeit ist die Beschränkung einfacher ersichtlich, denn wir können nur auf eine Stelle zur Zeit blicken und diese deutlich wahrnehmen. Allerdings ist damit nicht unbedingt festgelegt, dass unsere Aufmerksamkeit ausschließlich auf diesen Punkt gerichtet ist. Die Spotlight-Metapher dient einer bildlichen Beschreibung: Die Aufmerksamkeit wird wie ein Scheinwerfer auf verschiedene Bereiche gelenkt und je nachdem, wie der Scheinwerfer eingestellt ist, wird ein Punkt sehr hell erleuchtet (scharf gesehen) oder ein größerer Bereich in Augenschein genommen (zu Lasten der Sehschärfe) (vgl. J. R. Anderson, 2007, S. 100). Ungeübte Prozesse benötigen stärkere Aufmerksamkeit als automatisierte Prozesse. Beispielsweise ist das Lesen eines Wortes für funktionale Analphabeten und Analphabetinnen nicht automatisiert und erfordert eine größere Aufmerksamkeit als für Personen, die sich auf einem höheren Kompetenzniveau bezüglich des Lesens befinden.

Erfolgen zwei automatisierte Prozesse parallel, können trotz der Automatisierung Verzögerungen auftreten, weil zwei Informationen nicht gleichzeitig verarbeitet werden können. So zeigt z. B. der Stroop-Effekt, dass es Menschen schwer fällt, die Druckfarbe eines Wortes (z. B. blau) zu benennen, wenn es sich bei dem Wort um eine andere Farbe handelt (z. B. rot). Das Lesen eines Wortes

ist ein automatisierter Prozess, so dass es schwierig ist, ihn zu unterdrücken. Der automatisierte Prozess interferiert mit der Verarbeitung der anderen Informationen, die sich auf das Wort beziehen. Interferenzen können zudem auftreten, wenn zwei Reaktionen aufeinander folgen. Der Konflikt ist umso größer, je stärker die beiden Reaktionen auf die gleichen Verarbeitungsressourcen zurückgreifen (J. R. Anderson, 2007, S. 124).

Für die Gestaltung eines Systems lässt daraus ableiten, dass die Aufmerksamkeit durch akustische Signale oder visuelle Hervorhebungen gesteuert werden kann. Das kann zudem einen unterstützenden Effekt auf die Wahrnehmung haben.

2.6.3.2 Wahrnehmung

Die Wahrnehmung bezeichnet den Prozess und das Ergebnis der Informationsverarbeitung und steht dabei im Zusammenhang mit der Aufmerksamkeit: Die wahrnehmende Person muss entscheiden, worauf sie Aufmerksamkeit richtet. Dafür müssen die sensorischen Systeme erkennen, um was es sich in der Außenwelt handelt. Übertragen auf Computer und Interfaces ist die Frage, wie das Gehirn Symbole und Darstellungen auf dem Computerbildschirm erkennt und identifiziert. Bezüglich der Wahrnehmung ist grundlegend zwischen dem fovealen und dem peripheren Sehen zu unterscheiden (J. R. Anderson, 2007, S. 51). Bei Ersterem handelt es sich um die Reizaufnahme über einen kleinen zentralen Bereich der Netzhaut in dem scharf gesehen wird. Das foveale Sehen ist beispielsweise Voraussetzung für das Lesen. Beim peripheren Sehen werden umliegende Reize entdeckt, welche die Augenbewegung steuern.

Für die Gestaltung von Benutzer-/Benutzerinnenoberflächen spielen insbesondere die 1. *Objektwahrnehmung* und die 2. *visuelle Mustererkennung* eine Rolle:

1. *Objektwahrnehmung*: Visuelle Reize und Szenen werden mit Hilfe der sog. Gestaltprinzipien in Objekte gegliedert (J. R. Anderson, 2007, S. 60). Nach Wertheimer (1912) werden die visuellen Reize nach den vier Gestaltesetzen zu Objekten zusammengeführt.

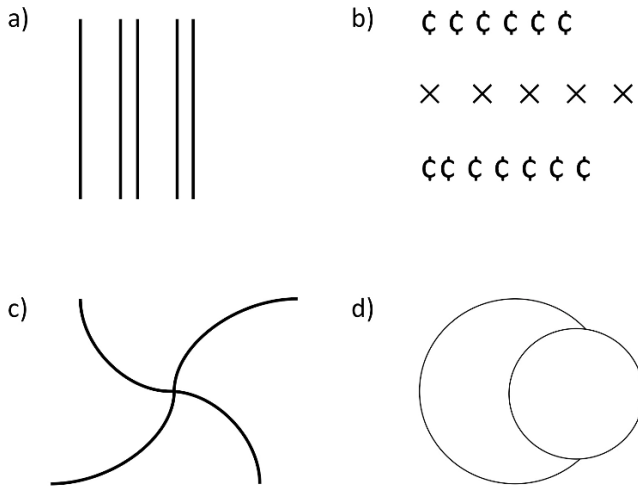


Abbildung 4: Grafische Darstellung der vier Gestaltgesetze in Anlehnung an Anderson (2007, S. 60)

- a) Das *Gesetz der Nähe* besagt, dass naheliegende Objekte als zusammengehörig empfunden werden.
- b) Tritt das *Gesetz der Ähnlichkeit* in Kraft werden ähnlich aussehende Objekte als Gruppe wahrgenommen.
- c) Beim *Gesetz des glatten Verlaufs* werden Linien so zusammengehörig wahrgenommen, dass sich ein glatter Verlauf anstatt von Ecken ergibt.
- d) Das *Gesetz der Geschlossenheit* bzw. der guten Gestalt drückt aus, dass Objekte als überdeckend wahrgenommen werden, wenn sich daraus eine geschlossene, klare Form ergibt, was z. B. eine räumliche Wahrnehmung hervorrufen kann.

Wertheimer beobachtete in Experimenten, in denen Striche auf verschiedene Weise angeordnet waren, dass Personen zwar nicht glauben, sie bewegen sich tatsächlich aufeinander zu doch sie sehen eine Tendenz der Striche zueinander „hin“. Je stärker die Aufmerksamkeit auf die beiden Objekte gerichtet war, desto stärker trat das Phänomen auf (Wertheimer, 1912, S. 226).

Die Gestaltgesetze können bei der Erstellung einer grafischen Oberfläche berücksichtigt werden, z. B. die Gesetze der Nähe und der Ähnlichkeit bei der Gestaltung eines Navigationsmenüs: Einerseits kann durch die aneinander liegende Anordnung der Menüpunkte bzw. -buttons zum Ausdruck gebracht werden, dass diese zu einer Kategorie gehören. Andererseits wird auch durch die

ähnlich Gestaltung der Schaltflächen der Eindruck erzeugt, dass die Objekte zu einer Gruppe (z. B. dem Webseitenmenü) gehören.

Neben der Gestaltung und Anordnung von Objekten kann auch der zeitliche Aspekt einen Einfluss auf die Wahrnehmung haben. Wenn Schaltflächen blinkend dargestellt werden, kann die Frequenz des Aufleuchtens tempoabhängig die Aufmerksamkeit auf sich ziehen. Dies trifft ebenfalls zu, wenn Zielreize durch Farbe, Größe oder Animation hervorgehoben werden. Wie bereits in Kap. 2.4 erläutert wurde, ist der *Extraneous Load* je nach Darstellung und Menge der Informationen ausgelastet. So kann auch für die Wahrnehmung gefolgert werden, dass für die Verarbeitung von Reizen je nach Menge und Darstellung die Gedächtniskapazität für die Wahrnehmung und Verarbeitung unterschiedlich stark beansprucht wird.

2. Visuelle Mustererkennung: Visuelle Informationen können in Objekte zergliedert werden, doch um die Welt sehen zu können, müssen die Objekte auch identifiziert werden. Dies geschieht mit der Mustererkennung. Sie liefert die Erklärung dafür, wie zum Beispiel die Darbietung des Buchstabens A auch als eine Ausprägung des Musters A erkannt wird. Eine Möglichkeit zur Erkennung ist der sogenannte Schablonenabgleich, bei dem ein Netzhautbild des Objektes an das Gehirn übermittelt wird, um es mit bereits gespeicherten Mustern zu vergleichen (J. R. Anderson, 2007, S. 61). Nach dieser Annahme versucht das Gehirn, das Bild eines Buchstabens mit jeder Schablone im Gehirn, die für Buchstaben vorhanden sind, abzugleichen und die Schablone mit der besten Übereinstimmung zu melden. Eine weitere Betrachtungsweise geht davon aus, dass die Mustererkennung auf einer Merkmalsanalyse beruht. Demnach wird jeder Reiz als Kombination elementarer Merkmale angesehen. Übertragen auf das Beispiel der Buchstabenerkennung sind dies die horizontalen, vertikalen, schrägen und/oder gekrümmten Linien der Buchstaben.

Bei der Wahrnehmung von Objekten werden einfache Teilobjekte (wie z. B. der lange Hals, vier Beine und ein Schwanz bei einer Giraffe) zu einem Gesamtobjekt zusammengeführt. Vertraute Objekte werden als Konfigurationen einzelner einfacher Komponenten zusammengefasst.

Schlussfolgernd sind Informationen entsprechend der Erkenntnisse zur Wahrnehmung so darzustellen, dass sie schnell erkannt bzw. erschlossen werden können. Starke Abweichungen von gängigen und geläufigen Mustern kann zu Irritationen, verlängerten Verarbeitungszeiten sowie einer starken Beanspruchung des *Extraneous Load* führen.

2.6.3.3 Wissensrepräsentation und Mentale Modelle

Nachdem Objekte wahrgenommen und erkannt (assimiliert) bzw. neue Objekte erschlossen (akkomodiert) wurden, werden die Informationen im Gehirn weiter

verarbeitet. Die Weiterverarbeitung von Informationen hängt davon ab, auf welche Weise sie im Gehirn repräsentiert sind (J. R. Anderson, 2007, S. 129). Grundlegend kann zwischen wahrnehmungsbasierter und bedeutungsbezogener Wissensrepräsentation unterschieden werden (J. R. Anderson, 2013, S. 97). Erstere bezieht sich auf direkt wahrgenommene Reize und ihre Repräsentationen im Gehirn. Bei der bedeutungsbezogenen Wissensrepräsentation werden die Informationen abstrahiert und bezüglich ihrer Bedeutung betrachtet. Der Nutzer/Die Nutzerin beurteilt auf der Basis des mentalen Modells die Folgen seiner/ihrer beabsichtigten Handlung. Mentale Modelle können durch Lernen und Erfahrungen permanent (weiter)entwickelt werden⁴⁶ (Niegemann, 2008, S. 284).

Im Kontext der HCI entwickeln Nutzer und Nutzerinnen über Eigenschaften und Funktionsweisen der Mensch-Computer-Interaktion und der Technologie mentale Modelle (Niegemann, 2008, S. 284). Auch bei der Verarbeitung visueller Informationen werden mentale Bilder verwendet (J. R. Anderson, 2007, S. 134). Sie werden bei Bedarf aus dem Gedächtnis aufgerufen. Wenn eine Person beispielsweise ein Computerprogramm aufruft, erzeugt die Person vor der eigentlichen Handlung (dem Anklicken des Programms) und dem eigentlichen Erscheinen des Programms – basierend auf ihrem bisherigen Wissen sowie bisherigen Erfahrungen mit Computerprogrammen und dessen Funktionsweise – eine mentale Vorstellung der Eigenschaften und Interaktionsmöglichkeiten. Mentale Modelle basieren also auf der subjektiven Wahrnehmung und nicht auf Fakten. Sie sind somit individuell – zwei Nutzer/Nutzerinnen können somit auch unterschiedliche mentale Modelle von einem Interface haben. Die Eignung eines mentalen Modells für die Arbeit mit einem Computerprogramm ist davon abhängig, wie systematisch und korrekt ein solches Modell entstanden ist (Herczeg, 2009, S. 52). Die Konstruktion mentaler Modelle von räumlichen Sachverhalten gelingt allerdings anhand von Texten weniger gut als anhand von Bildern (Brünken, Steinbacher, Schnotz & Leutner, 2001). Beispielsweise ist eines der größten Dilemmas, dass eine Lücke zwischen den gängigen mentalen Modellen von Interfacedesignern und den Nutzern/Nutzerinnen besteht. Begründet ist dies meist darin, dass Designer über zu viele Informationen verfügen und daher komplexere mentale Modelle entwickeln, die zu einer komplexeren Navigationsstruktur führen (Nielsen, 2010).

Da bildhaftes Material besser behalten werden kann als verbales (Brünken u. a., 2001; Sarodnick & Brau, 2011, S. 59), sind Symbole und Icons vermutlich besser geeignet, um einerseits mentale Modelle zu bilden als auch um mentale Modelle abzurufen. Um Informationen zu verarbeiten ist allerdings die duale

⁴⁶ Die Weiterentwicklung von mentalen Modellen wird als Assimilation, die Neuentwicklung von mentalen Modellen als Akkomodation bezeichnet (Graf, Zimbardo & Gerrig, 2007, S. 66). Diese Bezeichnungen gehen auf Jean Piaget zurück (vgl. z. B. Piaget, Aebli & Montada, 2003).

Repräsentation – visuell und verbal – effektiver. So erbringen Menschen bessere Gedächtnisleistungen, wenn Informationen visuell als auch verbal enkodiert werden (J. R. Anderson, 2007; Brünken u. a., 2001; Paivio, 1971).

Eine besondere Relevanz haben im Kontext der Nutzung von Programmen durch Personen mit niedriger Literalität Symbole und Icons, da diese ggf. die einzige Möglichkeit darstellen, mentale Modelle abzurufen. Wörter wären dafür weniger geeignet; erstens greifen nur mit einer geringen Wahrscheinlichkeit (in Abhängigkeit der Wortschwierigkeit) Automatismen (vgl. Kap. 2.6.3.3) und zweitens nimmt das Erfassen eines Wortes Zeit in Anspruch, erfordert hohe kognitive Ressourcen oder ist gar nicht erst möglich, so dass kein Sinn und somit auch kein mentales Modell abgerufen werden kann. Schlussfolgernd bieten Symbole das Potenzial, leichter mentale Modelle abzurufen aber auch falsch zu interpretieren. Werden beispielsweise erst mentale Modelle nach Aufrufen eines Symbols im Kontext eines Programmes gebildet, ist das mentale Modell mit dem programmspezifischen Symbol verknüpft und in seiner Funktionsweise richtig interpretiert. Werden Symbole aber uneinheitlich verwendet, kann das zu Konfusionen führen.

Kurz zusammengefasst ist die Wahrnehmung die Grundlage für die Aufnahme und Verarbeitung von Informationen. Symbole, Texte und Sprache müssen daher verständlich aufbereitet werden. Zudem sind Symbole so vereinfacht darzustellen, dass sie leicht im Gedächtnis bleiben und wiedererkannt werden können. Umgekehrt sind aber auch geeignete Eingabemöglichkeiten anzubieten. Für die Interaktion mit dem Medium Computer, z. B. über die Tastatur und Computermaus, sind daher auch physiologische Voraussetzungen relevant.

2.6.3.4 Physiologische Voraussetzungen

Zu den physiologischen Voraussetzungen bezüglich des Umgangs mit Computern zählen motorische sowie sensorische Fähigkeiten. Unter Motorik ist die mechanische Ausführungen von Handlungen zu verstehen. Motorische Fähigkeiten beziehen sich insbesondere auf den Tastsinn und die kinästhetische Sensitivität. Hierzu zählt auch die Bewegungskontrolle, welche die Handlung aufgrund von kognitiver und perceptiver Steuerung bezeichnet (Fisk, Rogers, Charness, Sharit & Czaja, 2009, S. 15). Bezüglich der HCI ist die Motorik die wesentliche Methode, um auf das System einzuwirken. Die motorischen Fähigkeiten können beispielsweise einen Einfluss auf die präzise Steuerung der Hände haben und somit den Umgang mit der Maus, die Platzierung des Mauszeigers auf den Schaltflächen und damit die Aktivierung von Schaltflächen (das Anklicken) erschweren. Auch kann durch motorische Einschränkung die Eingabe über die Tastatur beeinträchtigt sein. Im Laufe des Alterns nehmen diese Fähigkeiten

sukzessive ab⁴⁷. Beispielsweise reagieren ältere Menschen tendenziell langsamer (ältere Menschen reagieren durchschnittlich 1,5 bis zweimal langsamer als jüngere Erwachsene) und Bewegungen werden zunehmend unpräziser. Auch die Sensualität – das Empfindungsvermögen insbesondere bezüglich der Sinne Sehen, Hören, Schmecken, Riechen und Fühlen – nimmt tendenziell ab dem 60. Lebensjahr ab (Fisk u. a., 2009, S. 26). Allerdings spielen Riechen und Schmecken bei dem momentanen Stand der Technik im Usability-Kontext keine Rolle (wobei nicht auszuschließen ist, dass zukünftig auch diese Sinne via Computer angeregt werden können). Die Sinne Sehen und/oder Hören werden am Computer hingegen permanent angesprochen. Die Informationen mit Computern werden primär über den visuellen Kanal aufgenommen (z. B. bei Webseiten ohne auditive Unterstützung). Häufig werden aber auch gleichzeitig der visuelle und der auditive Kanal angesprochen (z. B. durch Filme). Die Wahrnehmung der Informationen hängt somit von der visuellen und auditiven Fähigkeit des Menschen ab, wobei insbesondere die Seh- und Hörfähigkeit mit zunehmendem Alter abnehmen (Sharit, Fisk, Rogers, Charness & Czaja, 2007, S. 17). Für die Usability und die UE sind diese heterogenen Voraussetzungen hinsichtlich der angenommenen Zielgruppe zu berücksichtigen. Ist beispielsweise davon auszugehen, dass in der Zielgruppe vermehrt Personen mit eingeschränkten motorischen Fähigkeiten vorhanden sind, sollten interaktive Schaltflächen größer im Verhältnis zur „Norm“ gestaltet werden. Zudem können Funktionen eingebunden werden, welche die Anpassung an individuelle Bedürfnisse ermöglichen, wie z. B. die Einstellungsmöglichkeiten von Lautstärke und Schriftgröße.

Werden zusammenfassend wieder die Darlegungen zu kognitiven und physiologischen Voraussetzungen mit den drei Einflussfaktoren Mensch, Technologie und Aufgabe in Beziehung gesetzt, so haben die menschlichen kognitiven und physiologischen Voraussetzungen einen Einfluss darauf, wie das Interface vom Menschen wahrgenommen wird, wie die Informationen verarbeitet werden und wie mit dem System interagiert wird. Diese Einflussfaktoren sind wiederum bei der Gestaltung der Technologie und der Aufgabe zu berücksichtigen und bilden die Grundlage für die Anforderungen an die Usability einer Software.

2.6.4 Allgemeine Usability-Anforderungen und Heuristiken

Es existieren unzählige Handbücher zum Thema Usability. Zudem existieren verschiedene Richtlinien (engl. *Guidelines*), Prinzipien, Grundsätze und Heuristiken, die dazu dienen sollen, einen hohen Grad an Usability sicherzustellen.

⁴⁷ In Deutschland ist über alle Alphastufen hinweg der Anteil funktionaler Analphabeten und Analphabetinnen bei den 50 bis 64jährigen Personen am größten (Grotlüschen & Riekmann, 2011, S. 7).

Neben dem internationalen Standard (EN ISO 9241) existieren weitere Qualitätskriterien und Heuristiken, die für den Großteil der Nutzer und Nutzerinnen Gültigkeit besitzen (vgl. z. B. Nielsen, 2012; Sarodnick & Brau, 2011; Shneiderman & Plaisant, 2005). Zudem bestehen Richtlinien für besondere Zielgruppen, z. B. ältere Personen (Fisk u. a., 2009) und Personen mit körperlichen und/oder geistigen Einschränkungen (Caldwell u. a., 2008). Der Unterschied zwischen Richtlinien und Heuristiken besteht darin, dass Richtlinien (im deutschen Sprachraum wird auch häufig das englische Wort *Guidelines* verwendet) stärker konkretisiert sind als Heuristiken.

Der Web Content Accessibility Guidelines (WCAG) (Richtlinien für barrierefreie Webinhalte) bestehen beispielsweise aus ca. 60 Richtlinien. Das United States Dept of Health and Human Services extrahierte auf der Grundlage von über 500 Richtlinien schließlich 209, die nach relativer Wichtigkeit und Aussagekraft beurteilt wurden. Diese sind in 17 Bereiche kategorisiert (z. B. UE, Texterscheingung, Accessibility und Seitenlayout) und bestehen wiederum aus mehreren konkret formulierten Richtlinien.

Die Richtlinien des DIN und des United States Dept of Health and Human Services sind in Abschnitte eingeteilt, die auch als grundlegende Prinzipien oder Heuristiken bezeichnet werden. Heuristiken sind leichter handhabbar, doch sind sie durch ihren Abstraktionsgrad von dem jeweiligen Nutzungskontext und der Interpretation der Experten und Expertinnen sowie deren Kompetenz abhängig (Hertzum, 2010, S. 571). Eine Heuristik ist die Lehre oder Wissenschaft von den Verfahren, um Probleme zu lösen mit Hilfe von methodischen Anleitungen und Anweisungen zur Gewinnung neuer Erkenntnisse. Sie besteht aus vereinfachten Annahmen, mit deren Hilfe ein Problem schneller gelöst werden kann als ohne Vereinfachung. Die Nachteile einer solchen Vereinfachung bestehen darin, dass eine Heuristik zwar schnell realisierbar ist, aber nicht mit Sicherheit zur korrekten Lösung führt (Gegensatz zu Lösungsalgorithmus) (Zimmer, 2004, S. 400).

Für die Überprüfung und Evaluation der Onlinetestumgebung für funktionale Analphabeten und Analphabetinnen werden Usability-Heuristiken als Evaluationsgrundlage herangezogen. Da bisher keine Erkenntnisse im Bereich der Usability-Anforderungen existieren und unter Berücksichtigung der vorhandenen Ressourcen für die Entwicklung und Evaluation ist eine Annäherung an eine erfüllte Usability über „grobe“ Heuristiken effizienter als über detaillierte Richtlinien. Im Folgenden werden verschiedene Usability-Heuristiken vorgestellt, kritisch reflektiert und zusammengefasst.

2.6.4.1 Usability-Heuristiken

Vorgestellt werden im Folgenden solche Heuristiken, die am stärksten etabliert (EN ISO 9241), viel zitiert (Nielsen, 2012), stark konkretisiert (Leavitt & Shneiderman, 2006) sind und besondere zielgruppenspezifische Charakteristika berücksichtigen, die auch auf funktionale Analphabeten und Analphabetinnen übertragbar sind (z. B. Fisk u. a., 2009). Die Heuristiken der DIN-Norm werden dabei am stärksten fokussiert, da sie sich am stärksten etabliert haben und im aktuellen Diskurs zur Usability-Forschung einen wesentlichen Referenz- und Bezugspunkt darstellen.

Die EN ISO 9241 ist der internationale Standard, welcher die Anforderungen an die Ergonomie der Mensch-System-Interaktionen beschreibt. Vorerst betitelte die International Standard Organisation (ISO) die Richtlinie mit „Ergonomische Anforderungen für die Bürotätigkeit mit Bildschirmgeräten“. Im Jahr 2006 wurde der Titel jedoch in „Ergonomie der Mensch-System-Interaktion“ umbenannt, um nicht nur auf Bürotätigkeit bezogen zu sein.

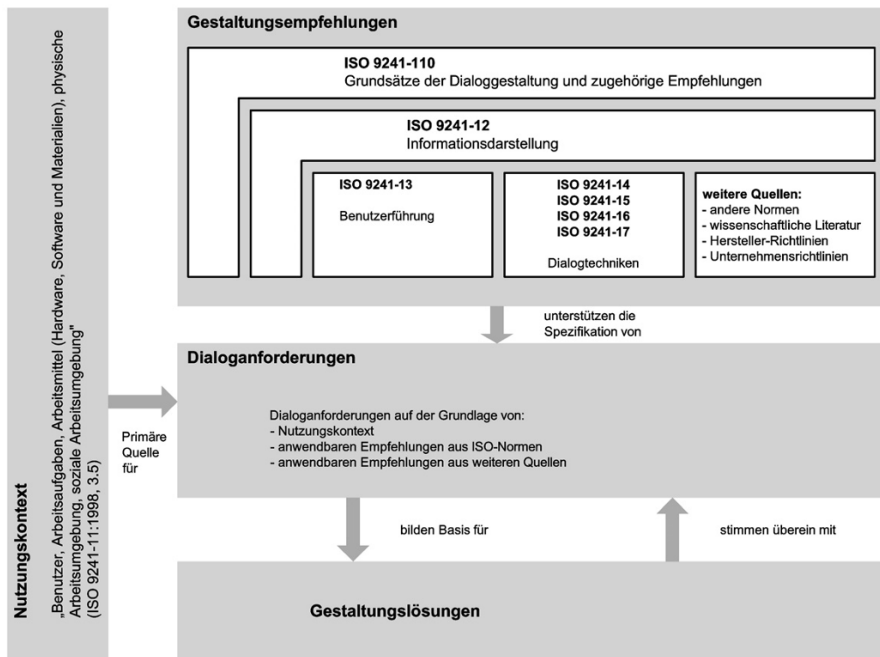


Abbildung 5: Gestaltungs-/Anwendungsrahmen für DIN EN ISO 9241 (DIN 2008, S. 18)

Die Norm besteht aus mehreren Teilen, wobei sich die Abschnitte 11 bis 17 und 110 dem Thema Software-Ergonomie widmen. Die Richtlinie 11 liefert die Definition zur Gebrauchstauglichkeit und stellt daher das „Kernstück“ der Definition von Usability dar. Sie beinhaltet vier Leitsätze: Leicht zu erlernen, intuitiv zu benutzen, geringe Fehlerrate sowie Zufriedenheit sicherstellen.

Der ursprüngliche Abschnitt 9241.10 „Grundsätze der Dialoggestaltung“ wurde ebenfalls 2006 umbenannt und trägt gegenwärtig die Kennziffer 9241.110. Dieser Abschnitt ist für die Entwicklung gebrauchstauglicher Webseiten essentiell. Die Grundsätze beziehen sich auf die Benutzer-/Benutzerinnenschnittstelle (engl. *User Interface*). Die Benutzer-/Benutzerinnenschnittstelle beschreibt die Bestandteile eines interaktiven Systems, welches die Steuerelemente zur Verfügung stellt, die für die Nutzer und Nutzerinnen notwendig sind, um eine bestimmte Aufgabe zu erledigen oder ein bestimmtes Ziel zu erreichen. Die Anforderungen an eine ausgeprägte Usability sind diesem Abschnitt zugeordnet.

Die Usability setzt sich nach DIN EN ISO 9241-110 zusammen aus der Effektivität, der Effizienz und der Zufriedenstellung einer Webseite (vgl. Kap. 2.6.1). Eine Webseite ist so zu gestalten, dass die Nutzer und Nutzerinnen ihr Ziel (Effektivität) mit möglichst wenig Aufwand (Effizienz) erreichen, um ein für sie zufriedenstellendes Ergebnis (Zufriedenheit) zu erzielen. Die Attribute Effektivität, Effizienz und Zufriedenheit werden im Folgenden kurz auf der Grundlage der DIN EN ISO 9241-110 erläutert.

Effektivität beschreibt den Grad, ob und wie genau ein Nutzer/eine Nutzerin sein/ihr Ziel erreicht. In Zusammenhang stehen damit die Gestaltung der Menü-, Navigations- und Orientierungsmittel, welche dies zulassen müssen. Die Effektivität entscheidet darüber, wie viel kognitiver Aufwand aufgebracht werden muss, um das Angebot formal handhaben zu können und wie viele kognitive Ressourcen demnach für das tiefere Verständnis und ggf. dem Wissenserwerb und Lernen verfügbar bleiben. Indikatoren für die Messung der Effektivität sind:

- ob das Ziel erreicht wurde
- welche/wie viele Informationen aufgerufen wurden
- welche/wie viele relevante Informationen nicht genutzt wurden.

Effizienz beschreibt das Verhältnis zwischen dem Aufwand und dem Ergebnis. Es werden beispielsweise Anstrengung, Aufrechterhaltung der Motivation und Zeit berücksichtigt. Ein Indikator zur Messung der Effizienz kann demnach die Zeit sein, die zur Erreichung des Ziels benötigt wird (vgl. auch Niegemann, 2008, S. 421; Rogers, Preece & Sharp, 2011, S. 14).

Zufriedenheit entsteht, wenn die Erwartungen des Nutzers/der Nutzerin und die Arbeitsprozesse ohne Einschränkungen erfüllt werden können. Die Zufriedenheit betrifft jegliche Aspekte der Gebrauchstauglichkeit. Zufriedenheit wird

gängigerweise über Fragebögen oder *Thinking Aloud*-Methoden (vgl. Kap. 2.7) erhoben (vgl. auch Niegemann, 2008, S. 421).

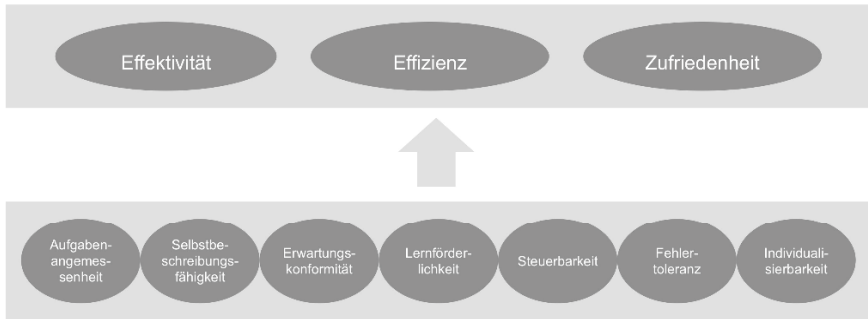


Abbildung 6: Grundsätze der Usability in Anlehnung an DIN EN ISO 9241-110 9241 (DIN 2008, S. 22)

Diese Attribute münden wiederum in sieben unterstützende Grundsätze der Dialoggestaltung (vgl. Abbildung 6): 1. Aufgabenangemessenheit, 2. Selbstbeschreibungsfähigkeit, 3. Erwartungskonformität, 4. Lernförderlichkeit, 5. Steuerbarkeit, 6. Fehlertoleranz und 7. Individualisierbarkeit (DIN Deutsches Institut für Normung e. V., 2008, S. 7):

1. „Ein interaktives System ist aufgabenangemessen, wenn es den Benutzer unterstützt, seine Arbeitsaufgabe zu erledigen, d. h., wenn Funktionalität und Dialog auf den charakteristischen Eigenschaften der Arbeitsaufgabe basieren, anstatt auf der zur Aufgabenerledigung eingesetzten Technologie“ (DIN Deutsches Institut für Normung e. V., 2008, S. 8).

Der Nutzer/die Nutzerin sollte auf einem direkten Weg das Ziel ohne zusätzliche belastende kognitive Anforderungen erreichen. Um diesen Anspruch zu erfüllen, sind im Vorfeld beispielsweise Aufgaben, Arbeits- und/oder Lernabläufe zu definieren, die einen potenziellen Nutzer/eine potenzielle Nutzerin ausführen möchte. Anschließend ist zu formulieren, wann eine Aufgabe erfüllt ist, um die Effektivität sicherzustellen. Letztlich ist der Weg zu definieren, der effizient – also mit geringem Einsatz von Zeit, Geduld, Gedächtnis- und Transferleistung – zu diesem Ziel führt.

2. *„Ein Dialog ist in dem Maße selbstbeschreibungsfähig, in dem für den Benutzer zu jeder Zeit offensichtlich ist, in welchem Dialog, an welcher Stelle im Dialog er sich befindet, welche Handlungen unternommen werden können und wie diese ausgeführt werden können“ (DIN Deutsches Institut für Normung e. V., 2008, S. 10).*

Dieser Definition nach muss sich der Benutzer/die Benutzerin zu jeder Zeit folgende drei Fragen beantworten können: Wo komme ich her? Wo bin ich? Wo kann ich von hier aus hin? Für die Beantwortung dieser Fragen ist die Besucherin/der Besucher in folgenden Aspekten zu unterstützen: a) Orientierung, b) Antizipierbarkeit, c) Feedback und d) Hilfe.

- a) Den Nutzer und Nutzerinnen müssen auf jeder Seite Orientierungspunkte angeboten werden, um zu erkennen, wo er/sie sich befindet und wie weit er/sie vom Ziel entfernt ist.
- b) Die Nutzer und Nutzerinnen müssen erkennen (antizipieren) können, wo die Navigationselemente hinführen. Erst damit wird die Seite steuer- und beherrschbar.
- c) Um den Nutzern und Nutzerinnen das Gefühl des Vertrauens zu geben, hat das System eine Rückmeldung darüber zu geben, ob Aktionen erfolgreich durchgeführt worden sind.
- d) Jedes System sollte Hilfestellungen anbieten, um auch unerfahrenen Nutzern und Nutzerinnen die Nutzung zu ermöglichen, insbesondere wenn die Seite und die auszuführenden Aktionen komplex sind.

Insgesamt ist das Ziel der Selbstbeschreibungsfähigkeit, dass Nutzer und Nutzerinnen in der Lage sind, die Software, die Internetseite etc. intuitiv richtig zu nutzen.

3. *„Ein Dialog ist erwartungskonform, wenn er den aus dem Nutzungskontext heraus vorhersehbaren Benutzerbelangen sowie allgemein anerkannten Konventionen entspricht“ (DIN Deutsches Institut für Normung e. V., 2008, S. 11).*

Das menschliche Verhalten ist geprägt durch erlernte Verhaltensmuster und Gewohnheiten. Je schneller Muster erkannt werden, desto schneller kann das Gehirn auch Zusammenhänge erfassen. Wenn Navigationselemente mit den gleichen Funktionen von Seite zu Seite variieren, können Nutzer und Nutzerin-

nen kein Muster speichern, welches ihnen die Navigation auf den folgenden Seiten erleichtern würde. Somit stünden auch weniger kognitive Ressourcen für den Inhalt zur Verfügung (vgl. Kap. 2.4 zur CLT). Je mehr Erfahrungen Nutzer und Nutzerinnen mit Software und Internetseiten haben, desto größer und präziser werden auch ihre Ansprüche bezüglich der Konformität mit ihren Erwartungen.

4. *„Ein Dialog ist lernförderlich, wenn er den Benutzer beim Erlernen der Nutzung des interaktiven Systems unterstützt und anleitet“ (DIN Deutsches Institut für Normung e. V., 2008, S. 12).*

Den Nutzern und Nutzerinnen ist die schnelle Aneignung von relevantem Wissen und Fertigkeiten zur Nutzung des Systems zu ermöglichen. Das Maß der Unterstützungsnotwendigkeit ist wiederum abhängig von den bisherigen Erfahrungen und gespeicherten mentalen Modellen. Es ist nicht möglich, jegliche individuelle Ausprägungen zu berücksichtigen, jedoch gilt, je seltener ein System genutzt oder eine Seite besucht wird, desto weniger ist ihm/ihr zuzumuten, die Bedienung zu erlernen (Rampl, 2007). Die Beanspruchung zusätzlicher kognitiver Ressourcen führt im Idealfall lediglich zur Verlangsamung der Interaktion, im schlimmsten Fall jedoch zur Unfähigkeit, den Dialog erfolgreich zu beenden. So komplex der Dialog auch sein mag, die Nutzer und Nutzerinnen verfügen über mentale Modelle, die ihnen mit geringer Transferleistung die Bedienung eines auf diese Muster angelegten Systems erlauben (Rampl, 2007). Nur in seltenen Fällen erfordert die Nutzung eines Systems das Erlernen neuer Navigationsarten. Das System sollte allerdings den Anspruch an einen logischen Aufbau und logische Abläufe sowie die Einbindung von Hilfeseiten/Hilfefunktion erfüllen. Die Lernförderlichkeit bestimmt somit die Hürde, die von den Nutzern und Nutzerinnen überwunden werden muss, um mit dem System zu arbeiten und ist daher ein wesentlicher Aspekt der Usability. Insbesondere ist sie von Bedeutung, wenn keine Schulungsmaßnahmen für die Nutzung vorgesehen sind bzw. das System eigenständig genutzt werden soll. Ein Maß zur Bestimmung der Lernförderlichkeit ist die Zeit, die Nutzer und Nutzerinnen brauchen, um ein bestimmtes Niveau im Umgang mit dem System zu erreichen (Niegemann, 2008, S. 424).

5. *„Ein Dialog ist steuerbar, wenn der Benutzer in der Lage ist, den Dialogablauf zu starten sowie seine Richtung und Geschwindigkeit zu beeinflussen, bis das Ziel erreicht ist“ (DIN*

Deutsches Institut für Normung e. V., 2008, S. 13).

Dazu zählen die Möglichkeiten, Medien zu nutzen oder auszuschalten, Alternativen zum Navigieren, einer Sicherstellung der korrekten Funktionsweise der Zurück-Schaltfläche, jederzeit die Startseite erreichen zu können sowie abubrechen. in Bezug auf Lernprogramme bedeutet Steuerbarkeit, dass Lernende jederzeit den Prozess unterbrechen und zu einem späteren Zeitpunkt fortsetzen können.

6. *„Ein Dialog ist fehlertolerant, wenn das beabsichtigte Arbeitsergebnis trotz erkennbar fehlerhafter Eingaben entweder mit keinem oder mit minimalem Korrekturaufwand seitens des Benutzers erreicht werden kann“ (DIN Deutsches Institut für Normung e. V., 2008, S. 14).*

Einerseits soll das System selbst wenig bzw. keine Fehler aufweisen, andererseits soll seitens des Systems verhindert werden, dass der Nutzer/die Nutzerin überhaupt erst Fehler macht. Falls dennoch Fehler gemacht werden, sollte eine schnelle Korrektur möglich sein. Fehler sind in diesem Fall Aktionen, die nicht zum gewünschten Erfolg führen (Niegemann, 2008, S. 423). Die Fehlerrate wird über das Aufsummieren der unerwünschten Systemreaktionen gemessen und kann zur Verarbeitung in folgende Kategorien eingeteilt werden: Vermeidbare Fehler (z. B. Programmierfehler oder Fehler, die mit Hilfe einer sorgfältigen Evaluation hätten verhindert werden können), bekannte, nicht vermeidbare Fehler (beispielsweise das Vertippen mit der Tastatur, das unbeabsichtigte Abschieken eines Formulars) und nicht antizipierbare Fehler (Fehler, die aufgrund unerwarteten Besucherverhaltens passieren). Bei der Entwicklung von Systemen sind schlussfolgernd Fehler, soweit möglich und antizipierbar, im Vorfeld zu vermeiden sowie bei nicht vermeidbaren Fehlern Wege aufzuzeigen, die den Nutzern und Nutzerinnen eine selbstständige Behebung erlauben.

7. *„Ein Dialog ist individualisierbar, wenn Benutzer die Mensch-System-Interaktion und die Darstellung von Informationen ändern können, um diese an ihre individuellen Fähigkeiten und Bedürfnisse anzupassen“ (DIN Deutsches Institut für Normung e. V., 2008, S. 15).*

Das Angebot soll sich an das individuelle Vorwissen und die Fähigkeiten anpassen lassen können. Bezüglich Webseiten sollte es möglich sein, individuelle Einstellungen vornehmen zu können, um einer heterogenen Nutzer-/Nutzerinnengruppe die Nutzung zu ermöglichen. Gängige Anpassungsmöglichkeiten beziehen sich auf die Sprache, die Schriftgröße und Vergrößerungsfunktion. Auch Filterfunktionen sind ein Merkmal von Individualisierung. Diese werden z. B. bei Einstufungstests verwendet, um anhand des Kenntnisstandes die optimale Aufgabenauswahl bereit zu stellen. Somit müssen nicht alle Inhalte genutzt bzw. bearbeitet werden, sondern nur diejenigen, die an den Kenntnisstand angepasst sind (Niegemann, 2008, S. 423).

Die Grundsätze stellen die wesentlichen Qualitätskriterien für die Gebrauchstauglichkeit bei der Entwicklung von Webseiten dar. Doch sind diese aufgrund ihres Abstraktionsgrades nicht immer einfach zu interpretieren und umzusetzen (vgl. auch Defizite allgemeiner Usability-Richtlinien in diesem Kap.). Die dargestellten Grundsätze sind jedoch nicht alle auf die computerbasierte Diagnostik zu übertragen. So kann die Steuerbarkeit auch absichtlich verhindert werden, um valide Ergebnisse zu erzielen: Während der Nutzung einer computerbasierten Diagnostik können ggf. Lerneffekte entstehen, so dass Nutzer und Nutzerinnen beispielsweise eine fehlerhafte Performanz korrigieren könnten. Auch die Individualisierbarkeit kann unter Umständen zu einem Reliabilitäts- und Validitätsproblem führen – je nach Abhängigkeit des Individualisierungsgrades und der Passung zwischen den Auswahlmöglichkeiten der computerbasierten Diagnostik und den Bedürfnissen oder dem Vorwissen der Nutzer und Nutzerinnen.

Neben der DIN Norm existieren weitere Qualitätskriterien, die je nach Nutzungskontext herangezogen werden können. Nielsen führt fünf Kriterien an, welche die Qualität der Usability definieren (Nielsen, 2012), die sich zum Teil mit den Normen der DIN überschneiden⁴⁸:

1. Erlernbarkeit (engl. *Learnability*): Wie einfach ist es für einen Nutzer/eine Nutzerin, bei der ersten Interaktion mit dem Interface, einfache Aufgaben auszuführen?
2. Effizienz (engl. *Efficiency*): Wie schnell können Nutzer und Nutzerinnen Aufgaben ausführen, sobald sie sich an das Design gewöhnt bzw. das Design „gelernt“ haben?
3. Erinnerbarkeit (engl. *Memorability*): Wie schnell können Nutzer und Nutzerinnen das erlernte Design reaktivieren, wenn sie über einen längeren Zeitraum die Seite bzw. die Umgebung nicht besucht haben?

⁴⁸ Auch die englischen Begriff finden im deutschen Sprachraum Verwendung, daher wird für jeden deutschen Begriff auch immer die englische Übersetzung aufgeführt.

4. Fehlertoleranz (engl. *Errors*): Wie viele Fehler können Nutzer und Nutzerinnen machen, wie schwerwiegend sind diese und wie einfach können sie diese Fehler beheben?
5. Zufriedenheit (engl. *Satisfaction*): Wie angenehm ist es die Seite zu nutzen?

In diese fünf Qualitätskriterien sind die Attribute Effizienz und Zufriedenheit der DIN EN ISO 9241-110 eingeschlossen, die Effizienz wird jedoch nicht explizit erwähnt.

Ein weiteres Qualitätskriterium, das Nielsen zusätzlich anführt aber nicht direkt der Qualität der Usability zuordnet, ist die Nützlichkeit, welche sich auf die Funktionalität bezieht. Hierbei geht es darum, in welchem Maß das Design die Absicht des Nutzers/der Nutzerin unterstützt. Nach Nielsen sind Usability und Utility gleich wichtig (vgl. auch Kap. 2.6.1 zur Begriffsan- und -einordnung).

Fisk u. a. (2009) formulieren Usability-Anforderungen für ältere Personen, die sich tendenziell durch eine niedrigere Computerkompetenz und eingeschränkte physische Leistungen (Sehfähigkeit und motorische Fähigkeiten) auszeichnen. Sie führen acht Prinzipien an, die sich ebenfalls zum Teil mit denen der DIN EN ISO 9241-110 und denen von Nielsen überschneiden (Fisk u. a., 2009, S. 75):

1. *Kompatibilität*: Das System sollte kompatibel mit den Erwartungen der Nutzer und Nutzerinnen sein.
2. *Konsistenz*: Die Platzierung von Symbolen und Schaltflächen sollte über die Screens hinweg gleich angeordnet sein und die gleichen Funktionen erfüllen.
3. *Fehlertoleranz*: Fehler von Nutzer und Nutzerinnen sollten von ihnen selbst leicht wieder behoben werden können.
4. *Feedback*: Die Resultate von Funktionen sollen klar sein und es sollte angezeigt werden, an welcher Stelle (der Bearbeitung) sie sich befinden.
5. *Individualisierbarkeit*: Nutzer und Nutzerinnen sollen die Möglichkeit haben, das System nach ihren Bedürfnissen anzupassen.
6. *Gedächtnis*: Das Gedächtnis der Nutzer und Nutzerinnen sollte nicht überladen sein und Unterstützungsfunktionen sollen angeboten werden.
7. *Struktur*: Die Interfaces sollen eine Struktur aufweisen, die den Nutzern und Nutzerinnen die Handhabung erleichtert.
8. *Workload*: Der Workload sollte so gering wie möglich gehalten werden, z. B. durch Hervorhebung von wichtigen Informationen.

Diese Guidelines sind zunächst abstrakt formuliert und werden hinsichtlich physischer Charakteristiken, Navigation, Informationsorganisation sowie konzeptioneller Ansprüche konkretisiert. Sie finden sich z. B. auch in der DIN EN ISO

9241-110 und in denen des United States Dept of Health and Human Services (2006) wieder. Allerdings liegt hier der Fokus darauf, spezielle Charakteristiken – wie z. B. eine Sehschwäche – zu berücksichtigen, indem beispielsweise die Schriftgröße anpassbar ist.

Molich und Nielsen (1990) haben versucht, die Komplexität der Regelbasis von Usability-Guidelines zu reduzieren und grundsätzliche Usability-Prinzipien als neun Heuristiken formuliert (Molich & Nielsen, 1990, S. 339):

1. Einfacher und natürlicher Dialog (engl. *Simple and natural dialogue*): Der Dialog am Bildschirm soll keine irrelevanten oder wenig nützlichen Informationen enthalten, denn jede irrelevante Information konkurriert mit den unverzichtbaren Aspekten des Dialogs. Die Informationen sollten in einer natürlichen und logischen Abfolge erscheinen.
2. Sprich die Sprache des Nutzers/der Nutzerin (engl. *Speak the user's language*): Der Dialog ist in verständlichen Wörtern und Sätzen zu erfolgen, die an die antizipierte Sprache potenzieller Nutzer und Nutzerinnen angepasst ist, anstatt einen systemorientierten Ausdruck bzw. Fachsprache der Computertechnologie zu verwenden.
3. Minimiere die Gedächtnislast der Nutzerin/des Nutzers (engl. *Minimize user memory load*): Das Arbeitsgedächtnis ist begrenzt (vgl. hierzu auch Kap. 2.4 zur CLT). Somit sollte von den Nutzern und Nutzerinnen nicht gefordert werden, sich an Informationen aus vorherigen Dialogen zu erinnern. Instruktionen zur Nutzung des Systems sollten klar sichtbar oder abrufbar an solchen Stellen sein, an denen sie auch gebraucht werden. Zudem sind Instruktionen in leicht verständlichem Ausdruck zu formulieren.
4. Sei konsistent (engl. *Be consistent*): Nutzer und Nutzerinnen sollten sich nicht fragen müssen, ob unterschiedliche Wörter, Situationen oder mögliche Aktionen das gleiche bedeuten. Eine bestimmte immer wiederkehrende Aktion sollte auch immer wieder durch die gleiche Aktion erreichbar sein (z. B. durch Anklicken einer bestimmten Schaltfläche).
5. Biete Feedbackfunktionen an (engl. *Provide feedback*): Das System sollte der Nutzerin/die Nutzerin permanent darüber informieren, wo er/sie sich befindet und welche Aktionen ausgeführt werden.
6. Platziere Möglichkeiten zum Verlassen des Systems klar und deutlich (engl. *Provide clearly marked exits*): Nutzer und Nutzerinnen sollten nicht in Bereichen „gefangen“ gehalten werden, indem keine Möglichkeit angeboten wird, das System oder die Seite zu verlassen. Es sollte immer die Möglichkeit eines „Notausgangs“ geben, der das schnelle Verlassen ohne das Durchlaufen eines längeren Dialogs ermöglicht.

7. Ermögliche die Nutzung von Shortcuts (engl. *Provide shortcuts*): Die Features, die ein System leicht erlernbar machen – wenig Eingabefelder und wortreiche Dialoge – sind oft beschwerlich für erfahrene Nutzer und Nutzerinnen. Intelligente Shortcuts sollten so in ein System integriert sein, dass das System sowohl erfahrene als auch unerfahrene Nutzer und Nutzerinnen anspricht.
8. Gute Fehlermeldungen (engl. *Good error messages*): Gute Fehlermeldungen sind defensiv, präzise und konstruktiv. Sie geben als Ursache immer das System an und kritisieren nicht den Nutzer/die Nutzerin.
9. Fehlerprävention (engl. *Error Prevention*): Noch besser als gute Fehlermeldungen ist es, Fehler und Probleme durch umsichtiges Design zu vermeiden.

Diese Heuristiken scheinen auf den ersten Blick offensichtlich zu sein, doch bei ihrer Überprüfung wird schnell deutlich, dass die Umsetzung nicht trivial ist. Molich und Nielsen (1990) stellten fest, dass Programmierer und Programmierinnen sowie Entwickler und Entwicklerinnen Schwierigkeiten in der Identifizierung potenzieller Probleme eines simplen Computerdialogs haben. Aus einer Evaluation eines Computerdialogs folgerten sie, dass überhaupt erst mal das Risiko fehlerhaften Designs von den Evaluierenden realisiert werden müsse. Darauf folgend sind Dialogprinzipien, wie die von ihnen formulierten, zu berücksichtigen um letztlich die Usability von vielen potenziellen Nutzern und Nutzerinnen überprüfen zu lassen. Werden Computersysteme von Personen entwickelt, die grundsätzliche Prinzipien der Dialoggestaltung verstehen und berücksichtigen, erreichen sie auch eine stärker ausgeprägte Usability. Die Ergebnisse der Evaluation weisen darauf hin, dass die Prinzipien weder selbstverständlich noch intuitiv bei der Entwicklung berücksichtigt werden (Molich & Nielsen, 1990, S. 342).

Stark daran angelehnt sind die „8 golden Rules“ nach Shneiderman und Plaisant (2005). Erweiternd werden von ihnen die Heuristiken mit *Design dialogue to yield closure*, *Permit easy reversal of actions* und *Support internal locus of control* angeführt (Shneiderman & Plaisant, 2005, S. 75). Die restlichen vier Regeln decken die neun Prinzipien von Molich und Nielsen ab. *Design dialogue to yield closure* beschreibt die Anforderung, Handlungssequenzen zu gruppieren in Anfang, Mitte und Ende. Eine Rückmeldung am Ende einer Handlungssequenz gibt dem Nutzer/der Nutzerin das Gefühl der Zufriedenheit, etwas abgeschlossen zu haben, ein Gefühl der Entlastung bzw. der Ablösung durch eine darauf folgende Handlungssequenz. *Permit easy reversal of actions* erhebt den Anspruch, dass Aktionen rückgängig zu machen sind. So wird Angst vor Fehlern aus der Perspektive der Nutzenden verringert und unterstützt das Ausprobieren von ggf. noch nicht bekannten Optionen. Die Heuristik *Support internal locus of*

control fordert, dass die Nutzer und Nutzerinnen überzeugt sind, die Kontrolle über ihre Handlungen zu haben. Dies beinhaltet auch die Möglichkeit, notwendige Informationen abrufen und notwendige Handlungen ausführen zu können. Im Zuge der technischen Entwicklung, der Standardisierung (inklusive der Richtlinien für die Barrierefreiheit – vgl. Kap. 2.6.4.2) und kulturellen Entwicklung können Heuristiken überholt, nicht mehr zutreffend sein und an Relevanz verlieren.

So haben Sarodnick und Brau (2011) haben 21 Jahre nach Nielsen und Molich 12 alternative Heuristiken zu den Heuristiken von Molich und Nielsen (1990) formuliert. Diese berücksichtigen sowohl die Prinzipien der bisher geläufigen als auch neue technische und kulturelle Entwicklungen. Erweitert werden die neun Heuristiken von Nielsen und Molich um die Heuristiken der *Individualisierbarkeit*, *Joy of use* und *Interkulturelle Aspekte* (Sarodnick & Brau, 2011, S. 149 f). *Individualisierbarkeit* bedeutet, das Dialogsystem sollte sich den individuellen Präferenzen anpassen lassen, so lange es der Effektivität, Effizienz und Zufriedenstellung dient und nicht im Widerspruch zu notwendigen technischen oder sicherheitsrelevanten Begrenzungen steht. *Joy of Use* bezieht sich darauf, dass Arbeitsabläufe und grafische Gestaltung – bei notwendiger Konsistenz – Monotonie vermeiden und zeitgemäß wirken sollten. Zudem sind Metaphern adäquat auf den Nutzungskontext abzustimmen. *Interkulturelle Aspekte* sollten dahingehend berücksichtigt werden, dass das System auf einen definierten Nutzer-/Nutzerinnenkreis und dessen funktionale, organisatorische und nationale Kultur abgestimmt sein sollte.

In den vorgestellten Heuristiken teilweise „mitgedacht“ aber nicht explizit berücksichtigt sind Personen mit körperlichen und psychischen Einschränkungen. Die WCAG knüpfen an den bereits vorgestellten Heuristiken an, werden jedoch um Empfehlungen speziell für Personen mit Behinderungen erweitert.

2.6.4.2 Accessibility und die Web Content Accessibility Guidelines (WCAG)

In dem Kapitel zur computerbasierten Diagnostik wurden Beispiele für Diagnoseinstrumente genannt, die für einen bestimmten Anlass entwickelt wurden, z. B. Vorauswahl von potenziellen Bewerbenden oder Steigerung der Literalität. Doch auch alltägliche Handlungen werden zunehmend über das Internet vollzogen – anfangen vom Einkaufen bis hin zur Anmeldung bei Behörden. Auch Personen mit Behinderungen nutzen zunehmend das Internet. Personen mit Behinderungen umfassen Personen mit visuellen, auditiven, motorischen, sprachlichen, kognitiven, Sprach-, Lern- und neurologischen Einschränkungen. Wie in den Kapiteln zu physiologischen und kognitiven Voraussetzungen erläutert, können Einschränkungen in diesen Bereichen einen großen Einfluss darauf haben, wie ein Interface wahrgenommen wird und insbesondere, ob eine computerbasierte

Diagnostik valide Ergebnisse liefert. Somit wird ein barrierefreier Zugang zu Internetseiten immer wichtiger.

Die Barrierefreiheit ist, wie auch die Ansprüche an die Usability in einer ISO-Norm (40500:2012) (International Organization for Standardization & International Electrotechnical Commission, 2012) geregelt, den sog. *Web Content Accessibility Guidelines* (WCAG) 2.0 (vgl. auch Caldwell u. a., 2008). Diese sind auf die Zugänglichkeit interaktiver Systeme (einschließlich Software) bezogen und stellen eine Grundlage dar, Zugangsbarrieren für Personen mit physischen und psychischen Einschränkungen abzubauen. Diese beziehen sich auf Interaktionen im Internet und basieren auf den vier Prinzipien (1) wahrnehmbar, (2) bedienbar, (3) verständlich und (4) robust. Darunter aufgeschlüsselt sind zwölf Richtlinien. Die Prinzipien und die dazugehörigen Richtlinien werden im Folgenden auf der Grundlage der Angabe des World Wide Web Consortiums (W3C) kurz dargestellt (Caldwell u. a., 2008):

Prinzip 1: Wahrnehmbar – Informationen und Bestandteile der Benutzer-/Benutzerinnenschnittstelle müssen dem Benutzer/der Benutzerin so präsentiert werden, dass sie diese wahrnehmen können. Hierzu zählen Richtlinien zu Textalternativen, zeitbasierten Medien (wie z. B. Audio und Video), Anpassbarkeit und Unterscheidbarkeit (beispielsweise bezüglich Farben und Kontraste).

Prinzip 2: Bedienbar – Bestandteile der Benutzer-/Benutzerinnenschnittstelle und Navigation müssen bedienbar sein. Richtlinien zu diesem Prinzip beziehen sich auf die Zugänglichkeit durch die Tastatur, ausreichend Zeit, Vermeidbarkeit von (epileptischen) Anfällen (z. B. dürfen dargestellte Inhalte nicht mehr als dreimal in kurzen Abständen blitzen) und Navigierbarkeit.

Prinzip 3: Verständlich – Informationen und Bedienung der Benutzer-/Benutzerinnenschnittstelle müssen verständlich sein. Hierzu zählen Lesbarkeit, Vorhersehbarkeit und Hilfestellung bei der Eingabe.

Prinzip 4: Robust – Inhalte müssen robust genug sein, damit sie zuverlässig von einer großen Auswahl an Benutzer und Benutzerinnen, einschließlich assistierender Techniken, interpretiert werden können. Dem Prinzip der Robustheit ist die Richtlinie zur Kompatibilität zugeordnet. Die Richtlinien geben die inhaltlichen Ziele vor, auf die Autoren hinarbeiten sollten, um Inhalte für die Nutzer und Nutzerinnen mit verschiedenen Behinderungen barrierefreier zu gestalten. Auch wenn die Richtlinien zur barrierefreien Gestaltung ein weiteres Feld an Bedürfnissen abdecken, können nicht alle Arten, Ausprägungen und Kombinationen von Behinderungen adressiert werden. Die Richtlinien verbessern die Nutzbarkeit explizit z. B. für ältere Personen mit sich altersbedingt ändernden Fähigkeiten. Sie verbessern damit häufig aber auch die Gebrauchstauglichkeit für Nutzer und Nutzerinnen im Allgemeinen (Caldwell u. a., 2008). Werden die

Ansprüche an einen barrierefreien Zugang berücksichtigt, steigt durch die klaren und einfachen Strukturen die Gebrauchstauglichkeit, die Kompatibilität wird verbessert und die Ladezeiten werden verkürzt. Das Thema *Accessibility* (engl. für Barrierefreiheit, Erreichbarkeit, Zugänglichkeit) im Kontext computerbasierter Diagnostik zu berücksichtigen, da diese auch online und ohne weitere Unterstützung stattfinden kann. Zudem ist die Zielgruppe der in diesem Forschungsprojekt untersuchten online und anonym zugänglichen Online-Plattform otu.lea tendenziell wenig gewohnt mit dem Computer zu arbeiten und profitiert vermutlich ebenfalls von einer barrierefreien Gestaltung. Die Zielgruppe von funktionalen Analphabeten und Analphabetinnen ist allerdings nicht mit der Zielgruppe der WCAG gleichzusetzen. Es kann zwar davon ausgegangen werden, dass einige funktionale Analphabeten und Analphabetinnen neben literalen Defiziten gleichzeitig Lernbehinderungen aufweisen (empirische Untersuchungen gibt es darüber bisher nicht, diese Vermutung gründet sich auf Erfahrungsberichten von Kursleitenden in der Alphabetisierung), jedoch ist diese Gruppe nicht durch Lernbehinderung oder physische Einschränkungen zu charakterisieren. Bei der Entwicklung einer Internetplattform für tendenziell computerungewohnte ist es daher zwar ratsam, auf die barrierefreie Nutzung zu achten – insbesondere im Kontext von onlinebasierten Kompetenzmessungen, da der Einfluss der ICT-Literacy auf das Testergebnis so weit wie möglich auszuschließen ist, um eine möglichst hohe inhaltliche Validität zu erzielen – jedoch ist es nicht notwendig und auch nicht ratsam, jegliche Richtlinien der WCAG zu berücksichtigen. Es ist zu vermuten, dass Anpassungsmöglichkeiten, die speziell für Personen mit Einschränkungen entwickelt wurden, auf eine geringe Akzeptanz stoßen und ggf. – aufgrund der empfundenen Gleichsetzung mit physisch oder psychisch eingeschränkten Personen – zum Abbruch führen würde.

Im Vergleich zu den Usability-Heuristiken wird hier deutlich, dass der Fokus stark auf die Nutzer und Nutzerinnen gerichtet ist und mit diesen Richtlinien der Anspruch erhoben wird, möglichst viele Personen mit unterschiedlichen Voraussetzungen den Zugang und die Nutzung zu ermöglichen. Gleichzeitig hat das Erfüllen bzw. Nicht-Erfüllen dieser Richtlinien einen Einfluss auf die Akzeptanz der potenziellen Nutzer und Nutzerinnen. Die DIN EN ISO 9241-110 versucht ebenfalls dem Anspruch gerecht zu werden, vielen Nutzern und Nutzerinnen den Zugang zu ermöglichen und schließt in ihren Richtlinien die barrierefreie Nutzung von Software zu einem großen Teil ein. Die WCAG liefern zusätzlich zu den Richtlinien ausführliche Hinweise, wie die Anforderungen inhaltlich und technisch erfüllt werden können. Es wird dabei allerdings weniger berücksichtigt, wie bei dem Entwicklungsprozess vorzugehen ist, um diesen Ansprüchen gerecht zu werden, zudem sind die Richtlinien und Prinzipien nicht explizit auf computerbasierte Diagnostik bezogen. Da aber auch in Institutionen

wie beispielsweise Schulen und Universitäten vermehrt computerbasierte Tests durchgeführt werden, bekommt die Barrierefreiheit auch hinsichtlich der computerbasierten Diagnostik eine stärkere Relevanz. Wird bei Tests in Universitäten und Schulen nicht auf eine barrierefreie Gestaltung geachtet, hat das womöglich Konsequenzen für die Schüler und Schülerinnen oder Studierende, da es ihnen unter Umständen nicht oder nur bedingt möglich ist aufgrund von Behinderung(en) an dem Test teilzunehmen und diesen erfolgreich abzuschließen. In England fand diese Überlegung ihren Anfang in dem *Disability Discrimination Act* (1995), in den USA mündeten die Bemühungen u. a. in dem *No Child Left Behind Act* 2001. Diese Anforderungen sind aber ebenfalls nicht auf elektronische Tests fokussiert. Erst 2002 wurde von der IMS Accessibility Project Group (IMS Global Learning Consortium) eine internationale Richtlinie für die Entwicklung von Lernapplikationen entwickelt (Crisp, 2007, S. 168). Diese Richtlinien und Standards führen allerdings bei Berücksichtigung nicht auch automatisch zu einem hohen Qualitätsstandard oder einer hohen Gebrauchstauglichkeit. Diese Lücke versucht Ball (2009) zu schließen, indem er den Fokus auf den Entwicklungsprozess barrierefreier E-Assessments richtet. Er formuliert vier Prinzipien für Entwickler und Entwicklerinnen, die es bei der Entwicklung eines E-Assessments zu berücksichtigen gilt (Ball, 2009):

- a) Prinzip des Antizipierens: Möglichst alle möglichen personalen Voraussetzungen für die Zugänglichkeit von E-Assessments müssen in dem Entwicklungsprozess berücksichtigt werden.
- b) Prinzip der Begründeten Anpassung: In vielen Entwicklungsprozessen begrenzt das Budget auch die Zeitspanne, in der Entwicklungen umgesetzt werden. Das hat oft eine reduzierte Umsetzung von Accessibility-Ansprüchen zur Folge. Es ist darauf zu achten, dass auf der Grundlage der materiellen Ressourcen klare Ziele formuliert werden.
- c) Prinzip der permanenten technischen Entwicklung: Im Zuge der technischen Entwicklungen werden auch neue Möglichkeiten für die Verbesserung der Zugänglichkeit entstehen. Daher sind die E-Assessments hinsichtlich ihrer Zugänglichkeit wiederholt zu überprüfen und die Usability ist entsprechend anzupassen.
- d) Prinzip der gemeinsamen Verantwortung: Die entwickelnde Institution hat eine klare Antidiskriminierungspolitik zu verfolgen. Zudem sollten sich alle Beteiligten aktiv für die Einhaltung dieser vier Prinzipien aussprechen.

Ball schlägt zudem praktische Schritte für die Entwicklung eines barrierefreien E-Assessments vor – angefangen von der Zusammenstellung des Entwickler-/Entwicklerinnen-Teams bis hin zum Akzeptanztest (Ball, 2009).

Um diesen Prinzipien und Anforderungen gerecht zu werden, bietet die momentane technische Entwicklung zahlreiche Möglichkeiten. Auf der technischen Ebene ist die Einbindung von Audio, Video und automatisierten Feedbackfunktionen möglich. Auf der Design-Ebene können Materialien „authentisch“ dargestellt werden. Die Diagnostik wird so mit multimedialen Elementen angereichert (vgl. Kap. 2.2.3 zum Thema Rich E-Assessment) und kann damit förderdiagnostischen Ansprüchen (vgl. Kap. 2.1.2) genügen.

2.6.4.3 Heuristiken – Grenzen der Anwendbarkeit

Auch wenn sich die Anwendung solcher Guidelines und Heuristiken etabliert hat, impliziert die Anwendung diverser Richtlinien nicht gleichzeitig, dass ein Interface auch gebrauchstauglich ist. Im Folgenden werden die Grenzen der Anwendbarkeit von Heuristiken (die auch auf Guidelines übertragbar sind) aufgezeigt.

Heuristiken sind immer kontextabhängig zu betrachten und anzuwenden. Vor der Anwendung ist zu prüfen, welche Heuristiken für den jeweiligen Zweck und die jeweilige Zielgruppe sinnvoll erscheinen und ob sie in Teilaspekten modifiziert werden müssen. Die Schwäche von Heuristiken liegt darin, dass sie meist nicht für eine spezielle Zielgruppe und Anwendungsumgebungen formuliert sind und meist unklar ist, wo die Grenzen der Interpretier- und Anwendbarkeit verlaufen (Bevan u. a., 1991, S. 654), woraus sich insbesondere drei Defizitbereiche von Guidelines ableiten lassen (Nielsen & Mack, 1994):

- Die Aussagekraft von Heuristiken hängt stark von dem Grad der Generalisierung ab. Je genereller eine Heuristik formuliert ist, desto weitläufiger sind auch Anwendungsfeld und Interpretationsspielraum. Um eine standardisierte Guideline zu interpretieren und anzuwenden, sind das Objekt, der Anwendungskontext, die potenzielle Zielgruppe sowie deren angenommene Eigenschaften hinsichtlich Präferenzen, Performanz und Lesefähigkeit zu definieren.
- Generelle Heuristiken können nicht „per se“ angewendet werden. Für die Anwendung bedarf es einer konkreten Definition des Nutzungskontextes: Die experimentellen Bedingungen, unter denen Richtlinien entwickelt und erprobt worden, sind nicht immer transparent. Somit können Heuristiken suggerieren, auf viele Anwendungskontexte übertragbar zu sein, ohne jedoch im Vorfeld einer Überprüfung unterzogen worden zu sein. Somit sind mit der Formulierung weitere, bisher nicht berücksichtigte oder getestete Anwendungskontexte ausgeschlossen, ohne dass die Grenzen abgesteckt werden. Falls nicht erwähnt wird, wie und wo die Heuristiken anzuwenden sind, bergen generell formulierte Heuristiken die Gefahr einer Missinterpretation.

- Das in den Richtlinien verwendete Vokabular stammt meist aus unterschiedlichen Disziplinen. Um eine erfolgreiche Anwendung zu gewährleisten und Fehlinterpretationen zu vermeiden, sind erfahrene Entwickler und Entwicklerinnen (aus möglichst unterschiedlichen Disziplinen) notwendig.

Wie auch die einzelnen Heuristiken immer kritisch hinsichtlich des Anwendungskontexts zu reflektieren sind, gilt auch für die eben genannten Defizite die Kontextabhängigkeit. Dennoch sind Richtlinien und die Verwendung von Checklisten in vielen Fällen unabdingbar, um formalen Standards gerecht zu werden. Die Heuristiken und Heuristiken sind je nach Domänenspezifität und Evaluationsgegenstand zu erweitern oder anzupassen. Dies geschieht immer unter der Prämisse, dass dadurch die Effektivität der Evaluation qualitativ und/oder quantitativ verbessert werden kann (Sarodnick & Brau, 2011, S. 148). Um generelle Heuristiken um domänen- oder produktspezifische Anforderungen zu erweitern, existieren nach Sarodnick und Brau zwei Möglichkeiten: Entweder durch die Einführung zusätzlicher Heuristiken oder durch die Erweiterung bzw. Modifikation bereits bestehender Heuristiken (Sarodnick & Brau, 2011, S. 148). Beispielsweise kann es bei einer computerbasierten Diagnostik sinnvoll sein, die Heuristik bezüglich der Möglichkeit, Eingaben zu korrigieren, umzuformulieren. So sollte ggf. die Möglichkeit bestehen, Eingaben unmittelbar nach der Eingabe zu korrigieren aber nicht unbedingt, nachdem weitere Aufgaben bearbeitet wurden.

Problematisch ist allerdings, dass universal formulierte Heuristiken suggerieren, bei ihrer Anwendung auch universelle Usability zu erreichen. Für die Anwender und Anwenderinnen von Heuristiken besteht eine Schwierigkeit darin, Heuristiken einzuordnen hinsichtlich ihrer Relevanz für das spezifische System. Eine Studie von Mosier und Smith (1986) weist darauf hin, dass oft weniger als die Hälfte der genutzten Heuristiken aufgrund von zu starker Spezifität, mangelnder Spezifität, verwirrend oder nicht zutreffend angewandt werden (Hertzum, 2010, S. 571). Die Herausforderung einer universellen Usability besteht in der Nutzer-/Nutzerinnen-Diversität, Wissenslücken und Technik-Variationen (Shneiderman, 2000). Studien weisen zudem darauf hin, dass existierende Heuristiken und weitere universell formulierte Usability-Prinzipien tendenziell entweder zu stark generalisiert oder zu vielzählig sind, um akribisch angewandt zu werden (Hertzum & Jacobsen, 2003; Mosier & Smith, 1986; P. Reed u. a., 1999). Die Richtlinien und Anforderungen an die Usability werden auf der Grundlage der Annahme formuliert, dass die Interfaces die größtmögliche Bandbreite an menschlicher Heterogenität abdecken. Shneiderman (2000, S. 85) formuliert den Anspruch an eine universelle Usability, die für 90% der Haushalte zutrifft. Durch diesen Anspruch ist es aber nicht möglich, spezielle

Charakteristika von potenziellen Nutzer-/Nutzerinnengruppen zu berücksichtigen. Z. B. sind mit der Zunahme der Nutzung der Technologie, beispielsweise um Formulare für Behörden und Ämter auszufüllen, auch spezielle Charakteristiken von Personengruppen bei der Entwicklung zu berücksichtigen, die nicht „dem normalen Nutzer/der normalen Nutzerin“ entsprechen. Um dieser Herausforderung zu begegnen versucht Hertzum (2010) beispielsweise, verschiedene Arten der Usability zu benennen und voneinander abzugrenzen: universelle, situationelle, wahrgenommene, hedonistische, organisationale sowie kulturelle Usability. Diese Differenzierung hat sich bisher nicht durchgesetzt. Zudem deutet alleine die Benennung der unterschiedlichen Arten von Usability auf Teilaspekte der Usability hin, die in allgemein formulierten Heuristiken und Guidelines vertreten sind. Erfolgreich dahingegen sind die Ansätze, Usability-Guidelines für bestimmte Zielgruppen zu formulieren, z. B. für Ältere (vgl. z. B. Fisk u. a., 2009) und Personen mit kognitiven und physischen Einschränkungen. Damit ist erneut darauf hingewiesen, dass Heuristiken unter Berücksichtigung des Nutzungskontextes und der potenziellen Zielgruppe auszuwählen, ggf. zu modifizieren und anzuwenden sind.

2.6.4.4 Zusammenfassung der allgemeinen Usability-Anforderungen und der WCAG

Für die Darstellung der oben genannten Ansprüche an eine ausgeprägte Usability wurden die Bezeichnungen („Attribute“, „Prinzipien“, „Guidelines“ und „Heuristiken“) von den jeweiligen Autoren und Autorinnen übernommen. Werden diese miteinander verglichen, wird deutlich, dass viele Überschneidungen bestehen. Um eine Übersicht über die grundlegenden Ansprüche an die Usability zu erhalten, wurden die oben vorgestellten Heuristiken einer genaueren vergleichenden Betrachtung unterzogen, um die Komplexität und Vielzahl der Anforderungen zu reduzieren. Ähnliche Heuristiken, wie z. B. „Reduziere die Auslastung des Kurzzeitgedächtnisses“ (Nielsen & Molich, 1990; Shneiderman & Plaisant, 2005), „Aufgabenangemessenheit“ (DIN) und „Workload“ (Fisk u. a., 2009) wurden unter der jeweils höheren Abstraktionsebene – in diesem Beispiel „Aufgabenangemessenheit“ – zusammengefasst. Daraus ergeben sich folgende Heuristiken:

1. Erlernbarkeit (DIN, 2006; Nielsen, 2006; Sarodnick & Brau, 2011): Der Nutzer/Die Nutzerin wird beim Erlernen des Systems angeleitet und unterstützt.
2. Konsistenz (DIN, 2006; Fisk u. a., 2009; Nielsen & Molich, 1990; Shneiderman & Plaisant, 2005): Die Gestaltung, Platzierung und Funktion von Symbolen und Schaltflächen ist über die einzelnen Interfaces hinweg konsistent.

3. Fehlertoleranz (Caldwell u. a., 2008; DIN Deutsches Institut für Normung e. V., 2008; Fisk u. a., 2009; Nielsen, 2012; Nielsen & Molich, 1990; Shneiderman & Plaisant, 2005): Meldet das System einen Fehler, müssen Art und Handlungszusammenhang enthalten sein. Handlungen müssen reversibel sein oder der Nutzer/die Nutzerin muss über irreversible Handlungen informiert werden. Zudem sollte das System von vorne herein so gestaltet sein, dass möglichst wenig Fehler passieren können.
4. Feedback (Fisk u. a., 2009; Molich & Nielsen, 1990; Shneiderman & Plaisant, 2005): Die Konsequenzen von Aktionen sind absehbar. Das System sollte dem Nutzer/der Nutzerin Transparenz über den Bearbeitungsstand und die Stelle, an der er/sie sich im System befindet, gewährleisten. Zudem sollte dem Nutzer/der Nutzerin eine Rückmeldung darüber gegeben werden, ob eine Aktion erfolgreich abgeschlossen wurde.
5. Individualisierbarkeit (Caldwell u. a., 2008; DIN Deutsches Institut für Normung e. V., 2008; Fisk u. a., 2009; Sarodnick & Brau, 2011): Das System soll für die Bedürfnisse der jeweiligen Nutzer und Nutzerinnen anpassbar sein (beispielsweise sollen Einstellungsmöglichkeiten für die Schriftgröße, die Darstellung von Grafiken und die Ein-/Ausschaltung von Multimedia vorhanden sein).
6. Selbstbeschreibungsfähigkeit (DIN, 2006; Nielsen & Molich, 1990; Sarodnick & Brau, 2011): Nutzer und Nutzerinnen sind in der Lage, das System intuitiv richtig zu nutzen und jegliche Dialogschritte nachzuvollziehen.
7. Wahrnehmungssteuerung (Caldwell u. a., 2008; Fisk u. a., 2009; Nielsen & Molich, 1990; Sarodnick & Brau, 2011): Die Interfaces sollen eine Struktur aufweisen, die leicht erschließbar ist und den Nutzern und Nutzerinnen die Handhabung erleichtern. Dies kann beispielsweise durch die Berücksichtigung der Gestaltgesetze (vgl. Kap. 2.6.3.2) erreicht werden. Das Layout sollte minimalistisch gestaltet sein. Die Aufmerksamkeit sollte mit Hilfe gestalterischer Elemente, die ausreichend kontrastreich gestaltet sind, auf relevante Informationen gerichtet werden.
8. Aufgabenangemessenheit (DIN, 2006; Fisk u. a., 2009; Molich & Nielsen, 1990; Sarodnick & Brau, 2011; Shneiderman & Plaisant, 2005): Die Nutzer und Nutzerinnen werden darin unterstützt, ihre Aufgaben und Ziele effektiv und effizient zu erreichen, d. h. ohne das Arbeitsgedächtnis zusätzlich zu belasten.
9. Erwartungskonformität (DIN, 2006; Sarodnick & Brau, 2011): Die Gestaltung des Systems sollte möglichst den Erwartungen der Nutzer und

Nutzerinnen entsprechen und deren Erfahrungen sowie Vorwissen bezüglich der Systemnutzung berücksichtigen.

10. Steuerbarkeit (DIN, 2006; Nielsen & Molich, 1990; Sarodnick & Brau, 2011; Shneiderman & Plaisant, 2005): Nutzer und Nutzerinnen sollten selbst über Start, Ende, Unterbrechung und Geschwindigkeit der Nutzung bestimmen können.
11. Perspektivübernahme (Caldwell u. a., 2008; Nielsen & Molich, 1990; Shneiderman & Plaisant, 2005): Es sollte die Sprache möglichst vieler potenzieller Nutzer und Nutzerinnen gesprochen werden, so dass der Inhalt ein hohes Maß an Verständlichkeit aufweist. Im Vergleich zur Erwartungskonformität bezieht sich die Perspektivübernahme insbesondere auf die Verständlichkeit der Inhalte wohingegen die Erwartungskonformität auf Handlungen bezogen ist.
12. Prozessangemessenheit (Sarodnick & Brau, 2011): Das System sollte für die Erfüllung realer Aufgaben in typischen Einsatzfeldern optimiert sein.
13. Datensicherheit (Sarodnick & Brau, 2011): Das System sollte auch bei fehlerhaften Eingaben und auch unter hoher Ressourcenbelastung stabil und ohne Datenverlust arbeiten.
14. Shortcuts (Nielsen & Molich, 1990): Sowohl für erfahrene als auch unerfahrene Nutzer und Nutzerinnen sollte die Nutzung gängiger Shortcuts möglich sein.
15. Joy of use (Sarodnick & Brau, 2011): Ohne die Konsistenz zu beeinträchtigen, sollte Monotonie vermieden und zeitgemäße Gestaltelemente eingebunden werden. Metaphern sind adäquat auf den Nutzungskontext abzustimmen.
16. Interkulturelle Aspekte (Sarodnick & Brau, 2011): Die kulturellen Hintergründe potenzieller Nutzer und Nutzerinnen sollten berücksichtigt und das System sollte darauf abgestimmt werden. Dies betrifft somit auch z. B. die Gestaltung des Storyboards und das Identifikationspotenzial mit dessen Charakteren.

Die Prinzipien der WCAG sind den vorgeschlagenen Heuristiken nicht eindeutig zuzuordnen. Diese überschneiden sich mitunter mehreren Heuristiken (so wäre z. B. das Prinzip „Verständlichkeit“ sowohl der Heuristik „Selbstbeschreibungsfähigkeit“ als auch der „Erwartungskonformität“ zuzuordnen, ebenso überschneiden sich Aspekte des Prinzips „Wahrnehmbar“ mit der Heuristik „Erwartungskonformität“), so dass die Prinzipien zwar in den Heuristiken mitunter berücksichtigt, nicht aber diesen zugeordnet sind.

Die Heuristiken stellen eine Zusammenfassung von Heuristiken für typische Nutzer und Nutzerinnen dar. In einem weiteren Schritt sind (vorerst auf theoreti-

scher Ebene) solche Heuristiken zu identifizieren, die ebenso für die in dieser Arbeit fokussierte Zielgruppe – funktionale Analphabeten und Analphabetinnen – geeignet sind.

2.6.5 Vorläufige Usability-Heuristiken für funktionale Analphabeten und Analphabetinnen

Die Relevanz der Heuristiken sowohl für funktionale Analphabeten und Analphabetinnen als auch vor dem Hintergrund des Produkts – der computerbasierten Diagnostik – zu bewerten. Im Folgenden wird erläutert und begründet, welche Heuristiken für die Entwicklung von otu.lea berücksichtigt wurden.

Ausgehend von der eben dargestellten Zusammenfassung wird die Heuristik „Shortcuts“ nicht berücksichtigt. Aufgrund der tendenziell niedrigen ICT-Literacy ist nicht zu erwarten, dass Teilnehmende Shortcuts verwenden, da es Irritationen hervorrufen könnte, wenn aus Versehen eine Tastaturkombination getätigt und ausgeführt wird, und der Effekt von den Teilnehmenden nicht nachvollzogen werden kann.

Die Individualisierbarkeit wird nur hinsichtlich gängiger Anpassungsmöglichkeiten (Größe und Lautstärke) umgesetzt. Weitere Anpassungsmöglichkeiten – wie z. B. die in dem Prinzip „Wahrnehmbar“ der WCAG formulierte Anpassung Kontraste – würden ggf. vom Inhalt ablenken oder die Teilnehmenden überfordern. Vor dem Hintergrund der theoretischen Überlegungen werden folgende Heuristiken für die Zielgruppe der funktionalen Analphabeten und Analphabetinnen als adäquat eingestuft und in weiteren Überlegungen zur empirischen Untersuchung berücksichtigt:

Vorläufige Usability-Heuristiken für funktionale Analphabeten und Analphabetinnen

1. Erlernbarkeit
2. Konsistenz
3. Feedback
4. Selbstbeschreibungsfähigkeit
5. Fehlertoleranz
6. Individualisierbarkeit
7. Wahrnehmungssteuerung
8. Aufgabenangemessenheit
9. Erwartungskonformität
10. Steuerbarkeit
11. Perspektivübernahme
12. Prozessangemessenheit
13. Datensicherheit
14. Joy of Use
15. Interkulturelle Aspekte

Tabelle 6: Vorläufige Usability-Heuristiken für funktionale Analphabeten und Analphabetinnen

Zusammenfassend ist festzuhalten, dass aufgrund der unterschiedlichen Heuristiken und der Abhängigkeit des Anwendungskontextes keine Sammlung oder Systematik von Qualitätskriterien oder Anforderungen an die Usability den Anspruch der Allgemeingültigkeit besitzen kann. Eine adäquate Usability für eine Online-Testumgebung ist gegeben, wenn die Gestaltung des Interfaces weder durch Handhabungs- noch durch Funktionsprobleme die Lösung der Aufgabe beeinflusst. Für die Gewährleistung dieses Anspruchs ist es unabdingbar, ausführliche iterative Evaluationen sowohl mit Usability-Experten und -Expertinnen als auch mit potenziellen Nutzern und Nutzerinnen durchzuführen. Auf der Grundlage von Evaluationen gilt es im empirischen Teil dieser Arbeit u. a., die Heuristiken zu überprüfen und ggf. anzupassen bzw. weitere Heuristiken aufzustellen.

2.7 Evaluation von Usability

Ziel der Usability-Evaluation ist, Schwierigkeiten und Probleme im Umgang mit dem System zu identifizieren. Usability-Evaluation bewertet nicht das System insgesamt – beispielsweise hinsichtlich seiner Relevanz im Bildungswesen – sondern danach, ob bestimmte Kriterien erfüllt sind. Ohne die Evaluation eines Produkts ist nicht festzustellen, inwieweit das Angebot den Bedürfnissen, Erwartungen und Fähigkeiten der potenziellen Zielgruppe entspricht. Zumeist verläuft die Usability-Evaluation – bzw. das Usability-Engineering, welches den Entwicklungs- und Modifikationsprozess bezeichnet – formativ. Das bedeutet, es gibt aufeinanderfolgende Designzyklen, mit der eine Datenbasis geschaffen wird, auf deren Grundlage Iterationen zur Verbesserung der Effektivität, der Effizienz und der Zufriedenheit vorgenommen werden (vgl. zur formativen Evaluation auch Kap. 2.1 und Kap. 3 zum DBR-Ansatz). Einzelne oder mehrere Phasen des Entwicklungsprozesses werden so lange in Zyklen durchlaufen bis die vorher definierten Entwicklungsziele erreicht wurden.

„Usability-Evaluation basiert auf dem Prinzip des nutzer-/nutzerinnenfreundlichen Designs“ (Niegemann, 2008, S. 426). Voraussetzung für die Evaluation ist die Erstellung eines Interface-Entwurfs, die Festlegung des Nutzungskontextes, die Erstellung einer Anforderungsanalyse sowie die Bestimmung der potenziellen Zielgruppe. Es ist festzustellen, was potenzielle Nutzer und Nutzerinnen als gebrauchstauglich und nutzer-/nutzerinnenfreundlich empfinden und wie sie mit dem System beabsichtigen zu arbeiten. Daraus werden des Weiteren notwendige Funktionen abgeleitet und definiert. Die Grundlage für Aussagen und Bewertungen der Usability sind die mittels empirischer Forschungsmethoden erhobenen Daten (Niegemann, 2008, S. 426). Für die Evaluation der Usability bestehen unterschiedliche Methoden, die nach unterschiedlichen Ansätzen systematisiert werden. Die Systematisierungen der Methoden unterscheiden sich einerseits in ihrem Grad der Differenzierung und andererseits hinsichtlich des Unterscheidungskriteriums.

Fisk et al. (2009) wählen einen geringen Differenzierungsgrad auf Grundlage des Kriteriums, ob Nutzer und Nutzerinnen einbezogen werden und differenzieren zwischen User Centered-Methoden („Thinking aloud“, Zielgruppenanalyse, Pilot-Testing, Feldbeobachtung, Fragebögen, Fokus-Groups) und Non User Centered-Methoden (Check-Listen, Heuristische Evaluation, Layout Analysen). Der Ansatz nach Sarodnick und Brau systematisiert anhand des Kriteriums Evaluationsmethoden und differenziert nach formal-analytischen Verfahren, Inspektionsmethoden, Usability-Tests und dem Einsatz von Fragebögen (Sarodnick & Brau, 2011). Der formal-analytische Ansatz wird im folgenden Kap. 2.7.1 näher erläutert. Der „reine“ Usability-Test wird im Kap. 2.7.5 erläutert. Inspektionsmethoden sind wiederum eine generelle Bezeichnung für ein Set an Methoden,

welche den Einsatz von Evaluierenden, die das Interface „inspizieren“, beinhaltet. Zu den Inspektionsmethoden zählen Heuristische Evaluationen und der Cognitive Walkthrough (vgl. auch Kap. 2.7.2). Als weitere Methoden, auf die aber nicht weiter eingegangen wird, werden „formal usability inspections“, „pluralsite walkthroughs“, „feature inspections“, „consistency inspection“ sowie „Standard inspection“ genannt (Sarodnick & Brau, 2011).

Nielsen (1994) zeigt vier grundlegende Wege auf, die Usability zu evaluieren, wobei die Unterscheidungskriterien die Erhebungsmethoden und -instrumente darstellen: automatisch (beispielsweise über das Loggen von Daten), empirisch (über die Testung des Produkts mit realen Nutzern und Nutzerinnen), formal (mit Hilfe von Modellen und formalen Anforderungen, auf deren Grundlage die Usability bewertet wird) sowie informell (basierend auf „Daumenregeln“ und der Fähigkeiten und Erfahrungen der Evaluatoren und Evaluatorinnen bzw. Entwickler und Entwicklerinnen). Nielsen bewertete allerdings zur Zeit dieser Systematisierung die automatischen Methoden als nicht umsetzbar und die formalen Methoden als zu aufwendig – insbesondere bei der Überprüfung von Produkten mit mehreren und/oder komplexen Benutzer-/Benutzerinnenschnittstellen (Nielsen, 1994, S. 413). Inzwischen sind automatisierte Methoden geläufig und werden vielfach eingesetzt.

Auch Niegemann u. a. (2008) differenzieren zwischen vier Kategorien der Evaluation: 1. dem formal-analytischen, 2. dem produktzentrierten, 3. dem interaktionszentrierten Messansatz sowie 4. dem benutzer-/benutzerinnenorientierten Messansatz (Niegemann, 2008, S. 430). Der Ansatz geht auf Bevan u. a. zurück (1991). Das Kriterium ist hierbei der Untersuchungsgegenstand unter Berücksichtigung des Entwicklungsstadiums.

Dieser Differenzierungsansatz spiegelt die beiden grundsätzlichen Untersuchungsgegenstände dieser Arbeit systematisch wieder, indem die Zielgruppe der funktionalen Analphabeten und Analphabetinnen durch die Methode des benutzer-/benutzerinnenorientierten Messansatzes sowie das computerbasierte Diagnoseinstrument durch die anderen drei Messansätze fokussiert werden. Dieser Systematisierungsansatz wird daher im Folgenden genutzt, um die Methoden der Usability-Evaluation darzustellen. Auch dient er als Grundlage des weiteren Vorgehens.

Im Folgenden werden die vier Ansätze erläutert. Der Fokus liegt dabei auf dem produktzentrierten, dem interaktionszentrierten und dem benutzer-/benutzerinnenzentrierten Messansatz, da diese in der Praxis häufig Anwendung finden und auch zur Beantwortung der Forschungsfrage dieser Arbeit herangezogen werden.

2.7.1 *Formal -Analytischer Messansatz*

Der formal-analytische Ansatz wird insbesondere in frühen Phasen der Entwicklung eingesetzt und bewertet die Usability auf der Grundlage formaler Eigenschaften der potenziellen Nutzer und Nutzerinnen (Performanz sowie psychomomentale Leistungen) und des Produktes. Beispielsweise wird (nach dem sog. Keystroke-Level-Modell) im Vorfeld antizipiert, welche Handlungen notwendig sind, um bestimmte Aufgaben zu erfüllen (z. B. Eingabe in die Tastatur) und wie viel Zeit dafür benötigt wird. Demnach kann dazu auch das Antizipieren des Verhaltens von potenziellen Nutzern und Nutzerinnen gehören. Hierfür werden typische Charakteristiken der Zielgruppe identifiziert (beispielsweise über Fragebögen) und in die weitere Entwicklung und Evaluation einbezogen. Diese Angaben werden mit erhobenen Daten in Usability-Tests verglichen. Systematische Abweichungen sollen auf Usability-Schwächen hinweisen. Häufige Kritikpunkte sind die Beschränkung auf den Zeitfaktor sowie der fehlende Einbezug von den Usability-Attributen Effektivität und Zufriedenheit (Niegemann, 2008, S. 429).

2.7.2 *Produktzentrierter Messansatz*

Beim produktzentrierten Messansatz steht das Produkt im Vordergrund. Es werden ergonomische Eigenschaften der Anwendung gemessen und beurteilt. Diese können wiederum differenziert werden in Inspektionsmethoden sowie Fragebögen und Checklisten (vgl. z. B. Niegemann, 2008, S. 200; Nielsen & Loranger, 2008).

Shneiderman und Plaisant (2005) fassen produktzentrierte Messansätze unter dem Begriff Expert-Reviews zusammen. Expert-Reviews sind Begutachtungen der Umgebung durch Experten und Expertinnen. Hierzu zählen a) die Heuristische Evaluation, b) das Guidelines Review, c) die Consistency Inspection, d) der Cognitive Walkthrough sowie e) die Formal Usability Inspection. Die Methode der Heuristischen Evaluation wird im Folgenden ausführlicher im Vergleich zu den weiteren Methoden dargestellt, um im empirischen Teil dieser Arbeit auf Teilaspekte der Heuristischen Evaluation zurückgreifen zu können.

a) Heuristische Evaluation

Bei der Methode der Heuristischen Evaluation handelt es sich um Discount-Usability-Engineering mit qualitativem Charakter (Sarodnick & Brau, 2011, S. 144). Bei der Durchführung werden Usability-Prinzipien berücksichtigen, die sich beispielsweise auf die Einhaltung von Konsistenz beziehen.

Bereits 1985 formulierten Gould und Lewis die drei Prinzipien eines guten Designs für nützliche und leicht handhabbare Computersysteme:

„Early Focus on User Tasks, Empirical Measurement und Iterative Design“ (Gould, 1985, S. 300).

Diese drei Prinzipien finden sich in jeglichen Heuristischen Evaluationen wieder. Eine Heuristische Evaluation gehört zu den informellen analytischen Verfahren und Inspektionsmethoden. Hierbei beurteilen eine geringe Anzahl von Experten und Expertinnen anhand von Heuristiken die Usability des zu analysierenden Produktes (Niegemann, 2008, S. 434). Der Erfüllung typischer Guidelines folgen vielzählige zu befolgende Regeln und rufen bei Entwicklern und Entwicklerinnen eher einen einschüchternden Effekt hervor als dass sie ambitioniert befolgt werden. Daher wählen Entwickler und Entwicklerinnen oft den Weg, Heuristische Evaluationen auf der Grundlage ihrer eigenen für sie relevanten Heuristiken durchzuführen (Nielsen & Molich, 1990, S. 249). Eine Heuristische Evaluation wird durchgeführt, indem Experten und Expertinnen ein Interface beobachten und dieses bewerten. Die Bewertung wird meist sinnvollerweise anhand von Kriterien – Heuristiken – vorgenommen. Einerseits kann durch die Nutzung von Kriterien aus der Perspektive der Entwickler und Entwicklerinnen der sie interessierende Fokus gesetzt werden. Andererseits kann dadurch der Vergleich von Beobachtungen verschiedener Personen einfacher gewährleistet werden (Nielsen & Molich, 1990, S. 249). Viel zitierte und angewandte Heuristiken sind die nach Nielsen und Molich (1990) und die „8 golden rules“ nach Shneiderman und Plaisant (2005) (vgl. Kap. 2.6.4.1).

Einen möglichen Ablauf einer heuristischen Evaluation beschreibt Niegemann (2008): Anhand der ausgewählten und ggf. modifizierten oder erweiterten Heuristiken beurteilen die einzelnen Experten und Expertinnen separat die Usability der Anwendung, idealerweise in zwei Durchgängen. Zudem können die Heuristiken im Team bewertet werden. Vor der heuristischen Evaluation sind die Experten und Expertinnen mit den heuristischen Prinzipien vertraut zu machen. Finden zwei Durchläufe statt, konzentrieren sich die Experten und Expertinnen im ersten Durchlauf auf den Informationsablauf und die Funktionalitäten, im zweiten Durchgang auf die einzelnen Bedienelemente. Die Usability-Probleme werden dabei schriftlich festgehalten, indem genau beschrieben wird, welche der heuristischen Prinzipien wie verletzt wurden. Nachdem die Probleme vorerst in Einzelarbeit analysiert und anschließend in der Gruppe der Experten und Expertinnen diskutiert wurden, wird eine Gesamtliste der identifizierten Usability-Probleme erstellt. In dem darauf folgenden Schritt werden die Usability-Probleme nach ihrem Schweregrad durch die Experten und Expertinnen beurteilt. Problematisch ist hierbei, dass es für den Schweregrad keine einheitliche Definition gibt. Folgende Faktoren sollten in die Beurteilung der Usability-Probleme einfließen (Nielsen, 2014): 1. Auftretenshäufigkeiten, 2. Einfluss des

Problems auf die Erreichung des gesetzten Ziels bzw. der Aufgabenbewältigung, 3. Persistenz des Problems.

Sarodnick und Brau (2011) beschreiben eine weitere Art der heuristischen Evaluation: die Kooperative Heuristische Evaluation. Dabei erarbeitet ein unabhängiger Versuchsleiter/eine unabhängige Versuchsleiterin zusammen mit den Entwicklern und Entwicklerinnen realistische Anwendungsszenarien. Usability-Experten und -Expertinnen werden mit dem System vertraut und mit den Soll-Prozessen bekannt gemacht. Während der Evaluation bearbeitet ein Usability-Experte/eine Usability-Expertin die Anwendungsszenarien; die evaluierende Person soll dabei die Schritte kommentieren. Der Usability-Experte/Die Usability-Expertin stellt Verständnisfragen zu Handlungsabfolgen und zu realen Arbeitstätigkeiten. Die Beschreibung realer Arbeitsschritte soll dem Usability-Experten/der Usability-Expertin dabei helfen, die Sicht der Nutzer und Nutzerinnen einzunehmen. Bei der Durchführung einer heuristischen Evaluation sind nach Sarodnick und Brau (2011) drei Faktoren zu berücksichtigen:

1. Der Usability-Experte/Die Usability-Expertin ist gleichzeitig Evaluator/Evaluatorin, Interviewer/Interviewerin und Lernender/Lernende. Die Fähigkeiten, sich während der Gesprächsführung in die Prozesse hineinzuversetzen und entsprechend zu reagieren, haben einen starken Einfluss auf den Evaluationserfolg.
2. Die potenziellen Nutzer und Nutzerinnen müssen die Inhalte der Prozesse verständlich und strukturiert wiedergeben.
3. Pro Räumlichkeit sollte nur eine Evaluation zurzeit stattfinden, da die Evaluierendenpaare sich gegenseitig ablenken oder beeinflussen könnten.

Die Heuristische Evaluation kann unterschiedlich stark detailliert oder umfangreich stattfinden. Es ist möglich, eine Heuristische Evaluation mit sehr wenig Aufwand durchzuführen, je nachdem, wie viele Heuristiken anhand wie vieler Aufgaben überprüft werden sollen. Daher handelt es sich auch um eine Methode des Discount-Engineerings.

b) Guidelines Review

Hierbei werden von Experten und Expertinnen relevante Guidelines erweitert und/oder modifiziert, um auf der Grundlage dieser das Interface hinsichtlich der Konformität mit den relevanten Guidelines zu überprüfen. Im Vergleich zur heuristischen Evaluation wird hier weniger systematisch vorgegangen.

c) Consistency Inspection

Bei dieser Methode wird der Fokus auf die Konsistenz der Interfaces gelegt, insbesondere wird darauf geachtet, dass Symbole, Terminologie, Schriftart, Farbschemata, das Layout sowie Input- und Output-Formate konsistent dargestellt sind.

d) Cognitive Walkthrough

Bei einem Cognitive Walkthrough werden Aufgaben und Handlungen von Experten und Expertinnen simuliert, die von potenziellen Nutzern und Nutzerinnen durchgeführt werden. Hierzu zählt zudem, einen typischen Tagesablauf eines potenziellen Nutzers/einer potenziellen Nutzerin zu antizipieren, um die mögliche Nutzung der Webseiten/des Programms im Kontext der Lebenswelt des Nutzers/der Nutzerin betrachten zu können. Die Methode des Cognitive Walkthrough wurde entwickelt, um insbesondere Interfaces zu evaluieren, deren Umgang explorativ erlernt wird. Cognitive Walkthroughs können durch Einzelne durchgeführt werden, wobei es jedoch sinnvoll ist, Cognitive Walkthroughs von möglichst unterschiedlichen Experten und Expertinnen (Programmierern/Programmiererinnen, Designern/Designerinnen, Nutzer/Nutzerinnen) durchführen zu lassen.

e) Formal Usability Inspection

Experten und Expertinnen präsentieren weiteren Entwicklern und Entwicklerinnen das Interface, um Stärken und Schwächen zu diskutieren und ggf. Lösungen für Probleme zu finden. Diese Form der Evaluation ist personell aufwändig und bleibt auf einer theoretischen Ebene. Ferner gibt sie Personen mit wenig Erfahrung die Möglichkeit, ihr Wissen über Usability-Anforderungen zu erweitern.

Zusammenfassend ist der produktzentrierte Messansatz in jeglichen Entwicklungsstadien anwendbar. Dadurch, dass potenzielle Nutzer und Nutzerinnen für die Durchführung nicht zwingend notwendig sind, kann der produktzentrierte Messansatz mit wenig Aufwand und schnell umgesetzt werden. Der Nachteil liegt jedoch auch genau darin: ohne potenzielle Nutzer und Nutzerinnen ist das Risiko groß, relevante Bedürfnisse von Nutzern und Nutzerinnen nicht zu erkennen und somit nicht berücksichtigen zu können.

2.7.3 Interaktionszentrierter Messansatz

Der interaktionszentrierte Messansatz fokussiert die Interaktion zwischen Mensch und Computer, wobei die Messung der Performanz sowie die psychomentalen Leistungen im Fokus stehen. Hierfür werden Probanden und Probandinnen aufgefordert, mit dem System zu arbeiten und ggf. bestimmte Aufgaben zu bearbeiten. Das Szenario gleicht damit dem eines Usability-Tests (das sogenannte „reine“ Usability-Testing wird in Kap. 2.7.5 erläutert). Es können Beobachtungen, Screen-Recording (die Handlungen des Nutzers/der Nutzerin werden videographiert), Logfile Recording (z. B. Speicherung der Anzahl und Reihenfolge der Seitenaufrufe, Dauer der Nutzung) und/oder Eye-Tracking (Aufnahme der Augenbewegungen) durchgeführt sowie Beobachtungsprotokolle erstellt werden. Eine spezielle Form, die im frühen Entwicklungsstadium eines zu entwickelnden Systems angewendet werden kann, ist das Paper Prototyping.

Paper Prototyping ist eine Methode, um Benutzer-/Benutzerinnenoberflächen zu entwickeln, zu testen und zu verbessern. Snyder (2007) definiert Paper Prototyping als eine Form eines Usability-Tests, in dem repräsentative Nutzer und Nutzerinnen realistische Aufgaben mit einer Papierversion der Benutzer-/Benutzerinnenschnittstelle durchführen (Snyder, 2007, S. 4). Eine Person simuliert dabei das Verhalten des Produktes (z. B. legt die entsprechenden Seiten vor, die erscheinen wenn eine bestimmte Schaltfläche aktiviert wurde) und reagiert nur auf die Aktionen der Nutzer und Nutzerinnen. Die computersimulierende Person liefert keine weiteren Erläuterungen und kommuniziert nicht mit dem Nutzer/der Nutzerin. Im Vorfeld ist ein Aufgabensample auf Papierbasis vorzubereiten. Des Weiteren sind alle Optionen, die durch das Aktivieren von Schaltflächen folgen können, vorzubereiten. Während des Paper Prototypings bekommt der Nutzer/die Nutzerin eine Papiermaus in die Hand, mit der er/sie eine Computermouse simuliert. Simuliert sie das Anklicken einer Schaltfläche hat die computersimulierende Person das Interface vorzulegen, welches erscheinen würde, wenn das Produkt entwickelt und programmiert ist.

Zur Teilnahme am Paper Prototyping sind potenzielle Nutzer und Nutzerinnen auszuwählen, die über das Ziel des Experiments, die Rahmenbedingungen und die Spielregeln informiert werden. Zudem werden sie aufgefordert, während des Experiments laut zu denken, um Aufschluss über weitere Aspekte zu erhalten, die womöglich nicht bedacht wurden. Vor der Durchführung sind die Rollen zu definieren und zu verteilen (Computer simulierende Person, Testperson bzw. Nutzer/Nutzerin, Beobachter/Beobachterin, Unterstützer/Unterstützerin für die Bereitstellung der Materialien). Die Dokumentation erfolgt idealerweise neben der Beobachtung über eine Videoaufzeichnung. Im Anschluss an das Experiment kann ein Interview mit dem Proband/der Probandin durchgeführt werden, um Informationen über die Wahrnehmung, Erfahrungen und Emotionen zu erhalten (Wolf & Koppel, 2010, S. 225). Die Methode dient zur Erstellung von Interface-Entwürfen durch interdisziplinäre Design-Teams (Hornecker, 2004) ebenso wie zur schnellen Evaluation von Entwürfen (Hackos, 1998; Rogers u. a., 2011).

Der Fokus liegt zwar auf dem Verhalten des Nutzers/der Nutzerin bzw. des/der Lernenden, es werden jedoch weniger deren Urteile erfragt (wie bei den benutzer-/benutzerinnenorientierten Messansätzen – vgl. Kap. 2.7.4), sondern ihre Handlungen protokolliert und später analysiert. Stehen Wahrnehmung und das Erleben der Nutzer und Nutzerinnen im Mittelpunkt des Interesses, werden benutzer-/benutzerinnenorientierte Messmethoden eingesetzt.

2.7.4 Benutzer-/Benutzerinnenorientierte Messansatz

Der benutzer-/benutzerinnenorientierte Messansatz wird angewendet, wenn primär die subjektive Beurteilungen potenzieller Nutzer und Nutzerinnen erhoben

werden sollen. Methoden zu dessen Feststellung sind Thinking Aloud (Lewis, 1982), Videokonfrontationsmethode (Neal & Simons, 1984), Fokus Groups (Nielsen, 1997), Fragebögen und Checklisten, sowie Usability-Tests mit dem Fokus auf die Wahrnehmung der Nutzer und Nutzerinnen. Für die Wahrnehmung der Nutzer und Nutzerinnen – die UE (vgl. in Kap. 2.6.2) – existieren standardisierte Instrumente; z. B. der AttrakDiff (Hassenzahl, Burmester & Koller, 2003) und der User Experience Questionnaire (Laugwitz, Schrepp & Held, 2006). Eine weitere Methode ist die Durchführung von Interviews – bei Niegemann auch als Question Asking bezeichnet (Niegemann 2008, S. 242). Diese Methode wird näher erläutert, da diese im empirischen Teil dieser Arbeit Anwendung fand. Eine umfassende Erläuterung wird in Kap. 8.2.2.1 vorgenommen.

Interviews können dazu dienen, weitere Aspekte aus der Perspektive der Nutzer und Nutzerinnen zu erfragen und Aufschluss über deren Wahrnehmung, Empfindungen und Einschätzung des Produktes zu erhalten. Da die Interviewmethode primär offene Frageformate verwendet, bietet sie z. B. auch die Möglichkeit, Verbesserungsvorschläge aufzunehmen. Im Vorfeld sollte dabei entschieden werden, welche Form von Interview durchgeführt werden sollte. Je nach Zielgruppe können unterschiedliche Methoden sinnvoll sein. So kann bei einem leitfadengestützten Interview generell die Wahrnehmung oder das Nutzungsverhalten von Personen im Vordergrund stehen. Bei fokussierten Interviews steht ein Gegenstand im Fokus, der beiden Interviewteilnehmenden bekannt ist. Da bei Usability-Tests meistens Informationen über die Wahrnehmung eines bestimmten Produktes und ggf. bestimmter Aspekte des Produktes erfragt werden sollen, eignet sich insbesondere die Methode des fokussierten Interviews. Hierfür ist im Vorfeld ein Leitfaden zu entwickeln und zu erproben. Zudem ist sicherzustellen, dass die Interviewer und Interviewerinnen die Fragetechniken beherrschen, um die interviewte Person nicht zu beeinflussen, z. B. durch Suggestivfragen (vgl. z. B. Helfferich 2004 und Kap. 8.2.2.1).

2.7.5 Usability-Tests in Laboratories

Usability-Tests werden durchgeführt, wenn das System in einem Entwicklungsstadium existiert, welches die Bearbeitung von realen Aufgaben durch Nutzer und Nutzerinnen ermöglicht. Mit Usability-Tests können Daten akquiriert werden, die sowohl Aufschluss über die Interaktion als auch über die potenziellen Nutzer und Nutzerinnen sowie die Passung von Mensch und System liefern. Damit sind auch die Fähigkeiten des potenziellen Nutzers/der potenziellen Nutzerin impliziert (vgl. z. B. Bevan u. a., 1991). Je nach Fokus können sie daher sowohl dem interaktions- als auch dem benutzer-/benutzerinnenorientierten Messansatz zugeordnet werden. Liegt der Fokus auf der Performanz und dem

System bzw. der Interaktion mit dem System, ist der Usability-Test dem interaktionszentrierten Messansatz zuzuordnen. Wird die Wahrnehmung und das Erleben mit dem System fokussiert, wird diese Methode dem benutzer-/benutzerinnenorientierten Messansatz zugeordnet (Niegemann, 2008, S. 444). Bei Usability-Tests in Laboratories werden Evaluationen in kontrollierter Umgebung – in Laboren – durchgeführt, um statistisch signifikante Unterschiede bezüglich der optimalen Gestaltung der Benutzer-/Benutzerinnenoberfläche zu erhalten. Dies geschieht dadurch, dass mindestens zwei experimentelle Bedingungen geprüft und verglichen werden. Hierfür werden Aufgaben ausgewählt, die hinsichtlich bestimmter Bedingungen variieren und von potenziellen Nutzern und Nutzerinnen durchgeführt werden. Damit soll festgestellt werden, welche Änderungen vorzunehmen oder – sollten Hypothesen vorab formuliert worden sein – welche Hypothesen zu verifizieren bzw. zu falsifizieren sind.

Bezüglich der Erhebung der subjektiven Beurteilung ist zu berücksichtigen, dass Maskierungseffekte auftreten können (Bevan u. a., 1991). So sind Usability-Tests in kontrollierter Umgebung keine Abbildung der Realität und geben somit ggf. nur bedingt Aufschluss darüber, wie sich die Nutzer und Nutzerinnen tatsächlich verhalten würden.

Für eine möglichst umfassende Problemanalyse und -aufdeckung sind weniger Kenntnisse der technischen Hintergründe entscheidend, sondern vielmehr domänenspezifische Kenntnisse. Nach Nielsen (1992) ist die Problemanalyse effektiver, wenn die Evaluierenden sowohl Domänen- als auch Usability-Expertise aufweisen. Nielsen spricht ihnen eine durchschnittliche Fehlerrückmeldung von 60% zu. Wird die Evaluation von Personen durchgeführt, die jeweils nur in einem der Bereiche Kenntnisse vorweisen, werden lediglich 22% bis 41% der Fehler entdeckt (1992, S. 376).

2.7.6 *Planung und Durchführung von Usability-Evaluation*

Bereits 1985 formulierte Shackel fünf Grundsätze für den Designprozess der Usability: „User-Centered Design“ (Designer und Designerinnen müssen Kenntnisse über die potenzielle Nutzer-/Nutzerinnengruppe und deren Handlungsinteresse haben), „Participative Design“ (ein Gruppe von potenziellen Nutzern und Nutzerinnen sollte eng mit dem Design-Team zusammenarbeiten und bereits früh in den Entwicklungsprozess einbezogen werden), „Experimental Design“ (potenzielle Nutzer und Nutzerinnen testen Papierprototypen und Prototypen in frühen Entwicklungsstadien), „iterative Design“ (Schwierigkeiten, die in früheren Entwicklungsstadien identifiziert wurden müssen modifiziert und einer weiteren Prüfung unterzogen werden) und „User supportive design“ (Unterstützende Funktionen für Nutzer und Nutzerinnen sind in frühe Entwicklungsprozesse einzubeziehen und in frühen Experimenten zu überprüfen) (Shackel, 1985, S. 20).

ff). Um aussagekräftige Ergebnisse zu erhalten, ist neben den zu berücksichtigenden Grundsätzen eine detaillierte Planung der Test-Durchführung notwendig, die folgende Schritte beinhaltet (angelehnt an Niegemann, 2008, S. 446 ff):

- 1) *Detaillierte Planung*: Definition von Testzweck, Testgegenstand, Zielgruppe, Auswahl von Erhebungsmethode und -technik, Räumlichkeiten und Testinhalt – z. B. das Aufgabensample – sowie die Entscheidung darüber, welche Daten erhoben und wie sie ausgewertet werden sollen. Auch ist eine Zeit- und Kostenabschätzung vorzunehmen.
- 2) *Zusammenstellung des Testmaterials und der Testumgebung*: Neben dem Untersuchungsgegenstand sind weitere Testmaterialien vorzubereiten. Dies können z. B. Fragebögen, Interviewleitfäden und psychologische Tests sein, die Aufschluss über die Wahrnehmung, das Empfinden, Gefallen und die Kompetenz der Nutzer und Nutzerinnen geben. Nach der Auswahl der Aufgaben sind die Szenarien vorzubereiten, es sind Briefingmaterial, Ablaufübersicht und weitere relevante Informationen für die Teilnehmenden zusammenzustellen. Auch sollten die Instruktionen für die Testleiter schriftlich vorliegen. Wichtig für die Verwendung der Daten ist die zu unterschriebene Einverständniserklärung der Teilnehmenden.
- 3) *Durchführung der Usability-Tests*: Je nach ausgewählter Methoden bzw. Kombination von Methoden wird der Usability-Test durchgeführt. Jede teilnehmende Person bearbeitet hierfür ein im Vorfeld definiertes Aufgabensample. Hierfür können zusätzlich Aufzeichnungsmethoden wie Videoaufnahme, Screen-Recording, Mouse-tracking und/oder Eye-Tracking verwendet werden.
- 4) *Instruktion der Teilnehmer und Teilnehmerinnen*: Die Teilnehmenden müssen über den Ablauf, das Ziel des Usability-Tests und Umgangsmöglichkeiten bei Problemen instruiert werden. Auch ist ihnen die Möglichkeit einzuräumen, Fragen zu stellen und – falls vorgesehen – spezielle Methoden wie beispielsweise lautes Denken einzuüben.
- 5) *Datensammlung und Analyse*: Nach der Datenerhebung müssen die Daten systematisch aufbereitet werden. Protokolle und Beobachtungen müssen schriftlich ausgearbeitet und für die Auswertung vorbereitet werden.
- 6) *Berichtlegung und Zusammenstellung*: Im letzten Schritt werden Empfehlungen zur Modifikation des Produktes in einem Ergebnisbericht zusammengestellt. Dabei sollte die Darstellung der Modifikationsnotwendigkeiten im Vordergrund stehen. Zu dem Bericht zählen die Vorstellung und Erläuterung der angewendeten Methoden und Techniken so-

wie die Darstellung der Ergebnisse und eine Auflistung der Empfehlungen.

Im empirischen Teil dieser Arbeit sind die eben dargestellten Grundsätze sowie die Ablaufschritte dahingehend umzusetzen, dass im Sinne des Grundsatzes „User-Centered Design“ zielgruppenspezifische Charakteristika berücksichtigt (vgl. Kap. 2.3 zu zielgruppenspezifischen Voraussetzungen im Theorieteil) sowie zielgruppenunterstützende Funktionen entwickelt und überprüft werden. Entsprechend des „Participative Design“ sind potenzielle Nutzer und Nutzerinnen bereits in frühen Entwicklungsstadien einzubeziehen, indem von ihnen erste Prototypen getestet werden („Experimental Design“). Des Weiteren ist bei dem Entwicklungsprozess der computerbasierten Diagnostik für funktionale Analphabeten und Analphabetinnen iterativ („Iterative Design“) in mehreren Entwicklungszyklen vorzugehen, so dass der Ablauf einer Usability-Evaluation – wie eben in Anlehnung an Niegemann dargestellt – nicht „singulär-linear“ erfolgt, sondern in mehreren Zyklen, in denen die einzelnen Schritte wiederholt werden. Den Rahmen für das empirische Vorgehen bildet der DBR-Ansatz (vgl. Kap. 3). Die Grundsätze der Usability-Evaluation werden in diesen Rahmen systematisch integriert und miteinander verknüpft.

2.8 Zusammenfassung der theoretischen Herleitung

Ausgehend von dem Alphabetisierungsbedarf wurden die aktuelle Situation funktionaler Analphabeten und Analphabetinnen sowie die Herausforderungen in der Alphabetisierung erläutert. Zur Reduzierung des funktionalen Alphabetismus kann eine computerbasierte Diagnostik beitragen, die eine effiziente Kompetenzdiagnostik und die Ableitung individueller Fördermaßnahmen ermöglicht. Die technische Entwicklung bietet Möglichkeiten der automatisierten Auswertung von offenen Antwortformaten, wodurch z. B. über die Auswertung von Textpassagen die Schreibkompetenzen von funktionalen Analphabeten und Analphabetinnen diagnostiziert werden können. Ziel des Forschungsvorhabens ist die Entwicklung und Evaluation einer Online-Testumgebung für funktionale Analphabeten und Analphabetinnen mit Single Choice- und (halb-)offenen Antwortformaten. Bei der Zielgruppe der funktionalen Analphabeten und Analphabetinnen ist allerdings von einer niedrigen ICT-Literacy auszugehen. Folglich ist das computerbasierte Diagnoseinstrument entsprechend gebrauchstauglich zu gestalten. Hinweise für die Gestaltung der Testumgebung liefern die Erkenntnisse zur Usability sowie zur CLT und CTML. Die folgende Abbildung stellt die Wirkungszusammenhänge der für die Forschungsfrage relevanten Faktoren reduziert dar:

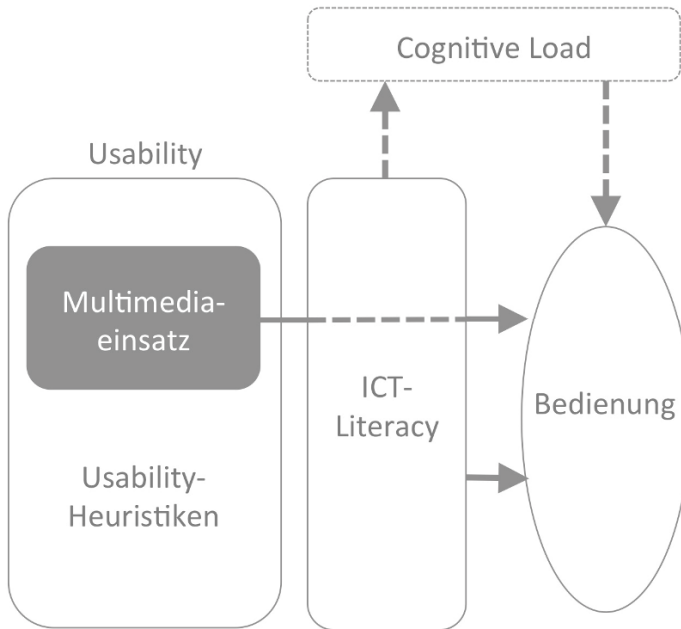


Abbildung 7: Wirkungsbeziehungen basierend auf der theoretischen Herleitung

Die durchgezogenen Linien verdeutlichen den vermuteten direkten Zusammenhang, die perforierten Pfeile weisen auf einen indirekten bzw. indirekt beobachtbaren Zusammenhang hin. Entsprechend der dargestellten theoretischen Basis ist die Usability eines Produktes durch die Effektivität, die Effizienz und Zufriedenheit charakterisiert. Zur Erreichung dieser Attribute dienen Usability-Guidelines und –Heuristiken. Der Einsatz von Multimedia kann zusätzlich zu einer ausgeprägten Usability beitragen. Im Kontext dieses Forschungsvorhabens wird die Usability über die Usability-Heuristiken und zudem den Einsatz von multimedialen Unterstützungsfunktionen evaluiert. Über das Bedienverhalten potenzieller Nutzer und Nutzerinnen können Rückschlüsse auf die Usability gezogen werden. Zu berücksichtigen ist dabei die tendenziell niedrige ICT-Literacy der funktionalen Analphabetinnen: Es ist davon auszugehen, dass die ICT-Literacy das Bedienverhalten moderiert und einen Einfluss auf die Beanspruchung des Cognitive Load – insbesondere des *Extraneous Load* – hat. Je nach Einsatz von Multimedia und der Ausprägung der Usability ist zu erwarten, dass die ICT-Literacy die Effekte, die schließlich über die Bedienung beobachtbar werden, beeinflusst: Beispielsweise ist bei einer niedrigen ICT-Literacy zu

vermuten, dass eine für den Nutzungszweck mangelnde Gestaltung einen stärkeren Einfluss auf das Bedienverhalten hat als Personen mit einer hohen ICT-Literacy. Als Zusatzexplanation dienen Erkenntnisse zur CLT: Beispielsweise ist bei einer niedrigen ICT-Literacy davon auszugehen, dass der *Extraneous Load* durch eine nicht nutzer-/nutzerinnenfreundliche Gestaltung stärker beansprucht ist als bei Personen mit einer hohen ICT-Literacy. Der Einsatz von Multimedia kann demnach die Reduktion der Beanspruchung des Cognitive Load unterstützen und die Bedienung der Online-Testumgebung beeinflussen.

Gemäß des DBR-Ansatzes (vgl. Kap. 3) wurde mit dieser theoretischen Herleitung der Alphabetisierungsbedarf und die mangelnden Ressourcen zur Reduzierung des Bedarfs als Problem in der Praxis identifiziert. Mit Hilfe aktueller Forschungserkenntnisse wurde das Forschungsvorhaben konkretisiert und hinsichtlich seiner praktischen sowie theoretischen Relevanz dargestellt. Damit ist die erste Phase des DBR-Prozesses abgeschlossen. Vor dem Hintergrund der bisher entfalteten theoretischen Basis (Computerbasierte Diagnostik, zielgruppenspezifische Voraussetzungen, CLT sowie der CTML und Usability) wurde die Online-Testumgebung otu.lea konzipiert und mit dem Design-Based Research-Ansatz evaluiert. In der folgenden Phase (Phase II) werden sowohl die Konzeption der Online-Testumgebung als auch die Konzeption des Forschungsrahmens erläutert. In der darauf folgenden Phase (Phase III) wird die Evaluation der Online-Testumgebung vorgestellt und es werden die in diesem Kapitel dargestellten Wirkungszusammenhänge im Kontext des Einsatzes der Online-Testumgebung für funktionale Analphabeten und Analphabetinnen untersucht. Abschließend erfolgt eine Reflexion des Prozesses (Phase IV).

Entwicklung einer Online-Diagnostik für die
Alphabetisierung

Eine Design-Based Research-Studie

Koppel, I.

2017, XIX, 399 S. 44 Abb., Softcover

ISBN: 978-3-658-15768-5