Fertigkeiten für die Lösung von kognitiven ICT-Aufgaben - Entwicklung und empirische Erprobung eines Erhebungs- und Validierungskonzepts

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt dem Fachbereich 05

Psychologie und Sportwissenschaften

der Johann Wolfgang Goethe-Universität

Frankfurt am Main

von

Lena Engelhardt

aus Heidelberg

Frankfurt am Main 2017

(D 30)

vom Fachbereich 05 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Dr. Winfried Banzer

Gutachter: Prof. Dr. Frank Goldhammer

apl. Prof. Dr. Karin Schermelleh-Engel

Prof. Dr. Klaus Kubinger

Datum der Disputation: 13. Juni 2018

INHALTSVERZEICHNIS

ZUSAMMENFASSUNG

Das Ziel der vorliegenden publikationsbasierten Dissertation liegt darin, ein Erhebungskonzept zu entwickeln, das es erlaubt, ICT-Skills – das heißt Fertigkeiten für das Lösen von Aufgaben in einer Informations- und Kommunikationstechnologie-Umgebung – theoretisch fundiert zu erheben sowie die Validität der intendierten Testwerteinterpretation empirisch zu untersuchen. Die Testwerte sollen als ICT-spezifische Fertigkeiten höherer Ordnung interpretiert werden.

Für die Erfassung von ICT-Skills kann auf keine lange Forschungstradition zurückgegriffen werden. Daher ist es das Ziel der *ersten Arbeit*, eine Rahmenkonzeption zur Messung von ICT-Skills zu erstellen. Dabei werden drei Ziele verfolgt: Erstens soll für die Itementwicklung spezifiziert werden, auf welchen generischen und ICT-spezifischen Fertigkeiten ICT-Skills basieren. Mithilfe etablierter psychologischer Theorien aus den relevanten Fertigkeitsbereichen werden kognitive Schwierigkeiten bei der Bewältigung von ICT-Aufgaben beschrieben, die als Grundlage für die Entwicklung der Items dienen. Zweitens werden für die Implementierung der Items Rationale für deren Erstellung in einer simulationsbasierten Umgebung formuliert, die es erlauben sollen, die intendierten kognitiven Prozesse realitätsnah in den Items abzubilden. Obgleich diese Arbeit einen konzeptionellen Fokus hat, besteht das dritte Ziel darin, die Rahmenkonzeption empirisch zu erproben, um zu beurteilen, ob die Rahmenkonzeption zur Itementwicklung und -implementierung geeignet war.

Aus der Rahmenkonzeption, die ein breites Spektrum relevanter ICT-Aufgaben für die Erfassung sowie eine simulationsbasierte Erhebung vorsieht, resultieren sehr heterogene Items. Deshalb unterscheiden sich ICT-Skills-Items von eher homogenen Itempools, wie sie typischerweise zur Erfassung von Konstrukten der psychologischen Leistungsdiagnostik, etwa zur Intelligenzdiagnostik, verwendet werden. Aus diesem Grund ist für die Konstruktvalidierung der Testwerteinterpretation, die das Ziel der zweiten und dritten Arbeit darstellt, zunächst konzeptionelle Forschungsarbeit nötig, um angemessene Validierungsstrategien für heterogene Items zu entwickeln. Diese in der zweiten und dritten Arbeit erforderlichen konzeptionellen Beiträge bedingen die Struktur dieses Rahmentextes, in dem zunächst die konzeptionellen Beiträge aller drei Arbeiten vorgestellt und anschließend alle empirischen Ergebnisse berichtet werden. Die konzeptionellen Entwicklungen für die Validierung der intendierten Interpretation der

Testwerte orientieren sich an Vorgehensweisen der psychologischen Leistungsdiagnostik, der nomothetischen Spanne und der Konstruktrepräsentation (vgl. Embretson, 1983). Mit diesen wird untersucht, inwiefern sich die zentralen Annahmen der Rahmenkonzeption aus der ersten Arbeit, nämlich die bei der Aufgabenlösung involvierten Fertigkeiten und kognitiven Prozesse, in den Testwerten widerspiegeln.

Das Ziel der *zweiten Arbeit* besteht darin, die nomothetische Spanne von ICT-Skills zu untersuchen und den postulierten Zusammenhang mit generischen und ICT-spezifischen Fertigkeiten empirisch zu untersuchen. Neben dem klassischen Ansatz, der Zusammenhänge über alle Items hinweg betrachtet, wird das Zusammenspiel verschiedener Fertigkeiten auch auf Itemebene analysiert. Darüber hinaus sollen potentielle Variationen in den Zusammenhängen über die sehr heterogenen Items durch Merkmale erklärt werden, welche für diese Heterogenität bezeichnend sind. Die empirischen Ergebnisse dienen – basierend auf den in der Rahmenkonzeption definierten Fertigkeiten – als Evidenzen für die Validität der Testwerteinterpretation.

Das Ziel der *dritten Arbeit* ist es, die Konstruktrepräsentation zu untersuchen, indem Evidenzen für die intendierten kognitiven Prozesse in der Itembearbeitung gesammelt werden. Klassischerweise werden in homogenen Itempools Itemmerkmale zwischen Items verglichen und wenn möglich quantifiziert, um die Schwierigkeit in Items zu beschreiben. Da die Items sehr heterogen sind, wurden zwei experimentelle Ansätze entwickelt, die diese kognitiven Prozesse in Itemvarianten verändern oder eliminieren. Die Auswirkungen dieser Manipulationen werden in Bezug auf die Itemschwierigkeit und den Zusammenhang mit anderen Konstrukten untersucht. Verändert werden die in der Rahmenkonzeption abgeleiteten schwierigkeitsdeterminierenden Merkmale, um zu untermauern, dass die ICT-Skills-Items ICT-spezifische Fertigkeiten erfordern. Eliminiert werden alle Merkmale die Fertigkeiten höherer Ordnung erfordern sollten. Mit diesen experimentellen Strategien können die zentralen Punkte der intendierten Testwerteinterpretation untersucht werden.

Neben den empirischen Ergebnissen zur Untermauerung der intendierten Testwerteinterpretation für den entwickelten ICT-Skills-Test ist der Erkenntnisgewinn dieser Arbeit auch in den konzeptionellen Beiträgen zu sehen. Mit diesen wurde exemplarisch gezeigt, wie ein Konstrukt wie ICT-Skills erfasst werden kann, indem man sich an den Vorgehensweisen der psychologischen Leistungsdiagnostik orientiert und dabei auf Annahmen kognitiver Prozesse zurückgreift.

1 EINLEITUNG

Kompetent mit Informations- und Kommunikationstechnologien (ICT) umgehen zu können, zählt zu den zentralen Fertigkeiten für das 21. Jahrhundert (Binkley et al., 2012) und zu den Schlüsselkompetenzen für lebenslanges Lernen (European Parliament and the Council, 2006). Die Bedeutung solcher Kompetenzen, welche im Folgenden ICT-Skills genannt werden, liegt vor allem darin, dass diese in allen Lebensbereichen benötigt werden und Menschen, welche nicht auf sie zurückgreifen können, Benachteiligungen erfahren. So werden ICT-Skills für Online-Geschäfte oder zur privaten Wohnungssuche, für einen Großteil der Arbeitsplätze im beruflichen Bereich sowie für das Erstellen schriftlicher Arbeiten oder Präsentationen im Bildungsbereich benötigt. Menschen, welche über geringere ICT-Skills verfügen, können dadurch Benachteiligungen erfahren, dass zum Beispiel günstige Wohnungsangebote gar nicht erst in einer Printversion zugänglich sind oder dass die Missachtung von Sicherheitsstandards bei Online-Geschäften zu einer unfreiwilligen Herausgabe persönlicher Daten führt. Im beruflichen Bereich ist es möglich, dass Menschen mit geringeren ICT-Skills wichtige Einstellungskriterien nicht erfüllen oder im Bildungsbereich Probleme bei der Informationsbeschaffung oder bei der Vermittlung von Informationen haben, zum Beispiel durch eine inadäquate Visualisierung. Oft reicht allein technisches Know-how nicht aus, um bestimmte Aufgaben zu lösen. So besteht bei einer Informationssuche im Internet die Herausforderung auch darin, Informationen hinsichtlich ihrer Qualität und Vertrauenswürdigkeit zu bewerten, weil das Internet nicht notwendigerweise Kontrollmechanismen bezüglich der Qualität bereithält (Rieh, 2002). Gerade jüngere Personen zeigen zwar weniger Probleme bei der Navigation und Orientierung im Netz, dafür aber Defizite beim Bewerten von Information (Eshet-Alkalai & Amichai-Hamburger, 2004; Eshet-Alkalai & Chajut, 2010; Lorenzen, 2001; van Deursen & van Dijk, 2009). Neben dem Alter hängt auch das Bildungsniveau mit ICT-Skills zusammen, zum Beispiel mit strategischen Fertigkeiten (van Deursen & van Dijk, 2009). Diese Fertigkeiten werden dafür benötigt, um die Möglichkeiten, welche ICT bieten, zielführend zu nutzen und Entscheidungen, zum Beispiel für ein bestimmtes Produkt, auf hochwertige Informationsquellen zu stützen. Im Fokus dieser Arbeit stehen solche ICT-spezifischen Fertigkeiten, welche dafür benötigt werden, um kognitive ICT-Aufgaben zu lösen.

Um die Fertigkeiten für den Umgang mit ICT erfassen zu können, wurden verschiedene Konzeptionen und Instrumente entwickelt (Ferrari, Punie & Redecker, 2012; Siddiq, Hatlevik, Olsen, Throndsen & Scherer, 2016). Die meisten Konzeptionen stimmen darin überein, dass für ein kompetentes Handeln neben rein technischem Wissen auch kognitive Fertigkeiten eine wichtige Rolle spielen. Hierzu wird jedoch eine große Bandbreite einzelner Fertigkeiten gezählt, unter anderem Lesen, Problemlösen, Rechnen, logisches Schlussfolgern, metakognitive Fertigkeiten oder kritisches Denken (Calvani, Cartelli, Fini & Ranieri, 2009, S. 186; International ICT Literacy Panel, 2002, S. 1). Diese werden auch als konventionelle Kenntnisse beschrieben („conventional literacies"; Fraillon, Schulz & Ainley, 2013, S.18). ICT-Skills benötigen also neben ICT-spezifischem Wissen auch generische Fertigkeiten, die nicht ausschließlich für ICT wichtig sind. Für die Operationalisierung von ICT-Skills wird in vielen Konzeptionen vorgeschlagen, die Breite des Gegenstandsbereichs in verschiedene kognitive Aufgaben aufzuteilen. Ein prominenter Ansatz ist der International ICT Literacy Panel (2002), in dem unterschieden wird, ob eine Aufgabe den Zugriff auf das Managen, Integrieren, Bewerten oder Erzeugen von Informationen erfordert. Diese Einteilung zeigt auch Überlappungen mit anderen Konzeptionen (z.B. Calvani et al., 2009; Eshet-Alkalai, 2004; Fraillon et al., 2013; National Higher Education ICT Initiative, 2003). Ferrari et al. (2012) beschreiben ein solches Vorgehen als aufgabenorientiert und grenzen es von Konzeptionen ab, die den Gegenstandsbereich nach Anwendungen wie Textverarbeitungs- oder Tabellenkalkulationssoftware unterteilen. Während ein anwendungsorientiertes Vorgehen eher verwendet wird, wenn die Zertifizierung von Fertigkeiten im Vordergrund steht, ist ein aufgabenorientiertes Vorgehen dann nützlich, wenn kognitive Fertigkeiten im Fokus stehen. Bei einer aufgabenorientierten Vorgehensweise bildet im Vergleich zu einem anwendungsorientierten Verfahren zum Beispiel die Bewertung von Informationen einen wichtigen Aspekt, unabhängig davon, ob diese in einem Browser oder einem E-Mail-Postfach stattfindet. Inwiefern kognitive Fertigkeiten wie Lesen oder Problemlösen nun aber zum Beispiel dafür benötigt werden, Informationen zu bewerten oder zu erstellen, wird in den Konzeptionen nicht spezifiziert. Eine solche theoretische Basis ist aber wichtig, um unter der Annahme schwierigkeitsdeterminierender Aufgabencharakteristiken Items entwickeln zu können (Kirsch, 2001). Genauso ist eine solche Basis wichtig, um im Zuge der Validierung der Testwerteinterpretation überprüfen zu können, ob die angenommenen kognitiven Prozesse, etwa hinsichtlich der Bewertung

von Informationen, tatsächlich in den Testwerten repräsentiert sind (Embretson, 1983; Kane, 2013).

Die vorliegende Dissertation hat zum Ziel, Fertigkeiten, die für das Bearbeiten von kognitiven ICT-Aufgaben benötigt werden, ökonomisch und in großem Maßstab wie in Large-Scale-Assessments erfassbar zu machen. Hierzu zählt neben einer theoretischen Rahmenkonzeption zur Entwicklung und Implementierung auch ein Konzept zur Validierung der Testwerteinterpretation. Diese beiden Zielsetzungen, die Entwicklung eines Erhebungs- und Validierungskonzeptes sowie deren empirische Anwendung, stellen die beiden Hauptziele dieser Arbeit dar.

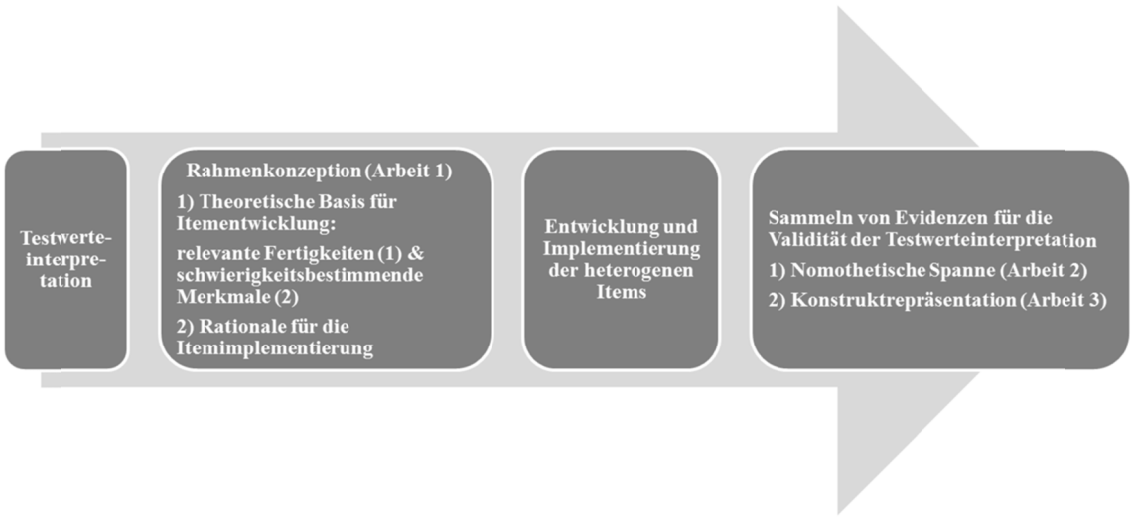


Abbildung 1. Die drei eigenständigen Arbeiten der vorliegenden publikationsbasierten Dissertation stellen aufeinander aufbauende Schritte im Zuge einer Testentwicklung dar.

Die erste Arbeit beschäftigt sich mit der Rahmenkonzeption (vgl. Abbildung 1), in der definiert wird, auf welchen Fertigkeiten ICT-Skills basieren und welche Merkmale in den Items schwierigkeitsbestimmend sein sollen. Obwohl die Arbeit einen stark konzeptionellen Fokus hat, wird anhand erster empirischer Ergebnisse untersucht, inwieweit die Rahmenkonzeption für die Erstellung der Items geeignet war. In der zweiten und dritten Arbeit werden dann Evidenzen für die Validität der Testwerteinterpretation gesammelt, wobei diese beiden Annahmen der ersten Arbeit jeweils empirisch überprüft werden. Hierbei werden die von Embretson (1983) beschriebenen Validierungsansätze der nomothetischen Spanne (Arbeit 2) und Konstruktrepräsentation (Arbeit 3) weiterentwickelt und durchgeführt. Die Untersuchung der nomothetischen Spanne basiert auf dem Zusammenhang mit anderen Variablen und untersucht die in der ersten Arbeit postulierten Zusammenhänge mit den relevanten Fertigkeiten. Die Untersu-

chung zur Konstruktrepräsentation beschäftigt sich mit den theoretischen Mechanismen, die den Itemantworten unterliegen, zum Beispiel mit dem Informationsprozess oder dem Wissen, das benötigt wird. Hierbei wird untersucht, ob die angenommenen kognitiven Prozesse, welche auf die in der ersten Arbeit beschriebenen schwierigkeitsdeterminierenden Aufgabencharakteristiken zurückzuführen sind, tatsächlich in der Itembearbeitung durchgeführt werden. Wie in der Abbildung 1 ersichtlich wird, bauen die drei Arbeiten aufeinander auf, da die beiden zentralen Aspekte der Rahmenkonzeption – die Definition von ICT-Skills in Bezug zu anderen Konstrukten und die Ableitung der schwierigkeitsbestimmenden Merkmale – jeweils separat in den Arbeiten zwei und drei im Zuge der Validierung betrachtet werden.

Mit ICT wird ein Gegenstandsbereich fokussiert, der sich in verschiedenen Punkten von solchen Konstrukten unterscheidet, die normalerweise in der psychologischen Leistungsdiagnostik untersucht werden, so zum Beispiel von Tests zur Erfassung von Intelligenz (z.B. BEFKI; Wilhelm, Schroeders & Schipolowski, 2014). Dies hat zur Folge, dass die Vorgehensweisen der psychologischen Leistungstestentwicklung nicht direkt übertragbar sind. An einigen Stellen im Prozess der Testentwicklung ist daher zunächst konzeptionelle Arbeit notwendig. Dies betrifft sowohl die Rahmenkonzeption als Grundlage für die Itementwicklung, die Itemimplementierung sowie das Vorgehen für die Betrachtung der Evidenzquellen auf Basis heterogener Items. Hieraus lassen sich zunächst konzeptionelle Ziele für die Testentwicklung und Validierung formulieren. Jede der drei Arbeiten lässt sich jeweils in zwei Teilbereiche aufgliedern: in einen konzeptionellen und einen empirischen Teil. Um den konzeptionellen Gedanken, der das Vorgehen in allen drei Arbeiten begründet, zusammenhängend zu beschreiben, werden die konzeptionellen Überlegungen im folgenden Kapitel (Kapitel 2) dargelegt. Diese Überlegungen bilden die Grundlage für die Herleitung der empirischen Fragestellungen (Kapitel 3). Die empirischen Ergebnisse basieren alle auf einer gemeinsamen Datenerhebung und werden nach einem gemeinsamen Methodenteil nacheinander ausgeführt (Kapitel 4). Die Ergebnisse aus den drei Arbeiten werden in Kapitel 5 integriert. Anschließend wird die Arbeit aus einer übergeordneten Perspektive diskutiert.

2 KONZEPTIONELLER RAHMEN

Der Chronologie von Abbildung 1 folgend soll zunächst der Gegenstandsbereich definiert und die intendierte Testwerteinterpretation dargelegt werden (Abschnitt 2.1), um anschließend die Rahmenkonzeption und die Rationale für die Itemimplementierung (Abschnitt 2.2) zu beschreiben. Die sich daraus ergebende Problematik eines heterogenen Itempools (Abschnitt 2.3) führt zu konzeptionellen Überlegungen zur Sammlung von Evidenzquellen für die Validierung der Testwerteinterpretation (Abschnitt 2.4).

2.1 Interpretation von ICT-Skills-Testwerten

Unter ICT-Skills werden in dieser Arbeit Fertigkeiten verstanden, die für das Verarbeiten und Produzieren von Informationen innerhalb einer Technologieumgebung benötigt werden. Hierbei werden alltägliche anfallende Aufgaben fokussiert, nicht aber professionelle Bereiche wie zum Beispiel der Informatikbereich. Mit diesem Fokus wird eine ähnliche Breite von Aufgaben angenommen, wie sie in modernen Bildungsvergleichsstudien wie PISA (Programme for International Student Assessment; OECD, 2014) oder PIAAC (Programme for the International Assessment of Adult Competencies; OECD, 2012) betrachtet werden. In diesen Studien wird untersucht, inwieweit solche Aufgaben gemeistert werden, die im alltäglichen Leben auftreten.

Alltagstypische Aufgaben können verschiedene Ebenen an Fertigkeiten erfordern. Eher basale ICT-Skills (z.B. Basic Computer Skills; Goldhammer, Naumann & Keßel, 2013) werden für Aufgaben benötigt, die durch basale Handlungen in einer ICT-Umgebung gelöst werden können, beispielsweise für das Weiterleiten einer E-Mail. Hingegen werden ICT-Fertigkeiten höherer Ordnung für Aufgaben benötigt, in denen zusätzlich Entscheidungen getroffen werden müssen, zum Beispiel darüber, ob eine E-Mail weitergeleitet werden soll oder nicht, nachdem diese bewertet wurde. Um zu beschreiben, wie diese Aufgabenmerkmale und erforderlichen Fertigkeiten zusammenhängen, wird eine Handlung im ICT-Kontext in Abbildung 2 schematisch vereinfacht dargestellt. Die Abbildung 2 basiert auf einem allgemeinen Modell der hierarchisch-sequentiellen Organisation einer Handlung nach Volpert (1982) und wurde für den ICT-Kontext im Rahmen dieser Dissertation angepasst. Eine Handlung wird hier durch verschiedene Handlungsebenen beschrieben. Auf der untersten Ebene liegen die einzelnen Interaktionen mit der ICT-Umgebung, die Lösungsschritte. Ein solcher Lösungsschritt kann das Drücken eines Buttons sein, um zum Beispiel eine E-Mail zu ver-

senden, oder das Eintippen einer E-Mail-Adresse. Eine oder mehrere dieser Lösungsschritte bilden eine Teilhandlung, zum Beispiel das Weiterleiten einer E-Mail, bestehend aus dem Drücken einzelner Buttons und dem Eintippen der E-Mail-Adresse. Um eine Entscheidung zu treffen, bedarf es einer oder mehrerer Teilhandlungen, zum Beispiel um Informationen für die Entscheidungsfindung einzuholen. Solche getroffenen Entscheidungen bedingen wiederum andere Teilhandlungen, beispielsweise das Versenden einer E-Mail an einen bestimmten Adressaten. Eine Aufgabe kann das Treffen einer oder mehrerer Entscheidungen beinhalten. Um im gewählten Beispiel zu bleiben, könnte die Aufgabe darin bestehen, eine bestimmte E-Mail weiterzuleiten (eine Entscheidung) oder auszuwählen, welche von fünf eingegangenen E-Mails weitergeleitet werden soll (mehrere Entscheidungen). Je nachdem, wie die Entscheidung getroffen wird, erfolgen weitere Teilhandlungen und Lösungsschritte.

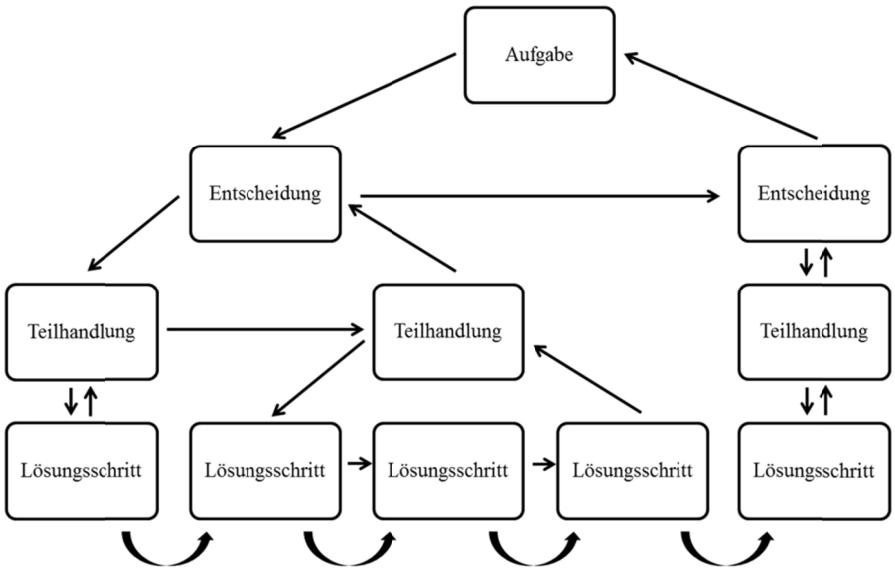


Abbildung 2. Schematische Darstellung einer Handlung in ICT-Umgebung, adaptiert nach Volpert (1982).

Das hier beschriebene Modell zeigt auch Parallelen zu der Konzeption von van Deursen und van Dijk (2011). Diese Autoren teilen die zur Bewältigung von Aufgaben im Internet erforderlichen Fertigkeiten in vier Ebenen ein. Die Fertigkeiten der ersten beiden Ebenen werden als relevant angesehen, um mit dem Internet als Medium umgehen zu können, und erfordern operationale und formale Fertigkeiten. Die beiden höheren Ebenen werden als inhaltsbezogen beschrieben und erfordern Informationskompetenz sowie strategische Fertigkeiten für das Internet. Die inhaltsbezogenen Fertigkeiten werden in dem Handlungsmodell für das Treffen der Entscheidungen benötigt, während

die Fertigkeiten der beiden unteren Ebenen für die Teilhandlungen und Lösungsschritte erforderlich sind.

Alltagstypische Aufgabenstellungen benötigen zumeist das Treffen von Entscheidungen. Aufgaben, welche solche Entscheidungen erfordern, gehen über reine Routine-Aufgaben hinaus, die allein durch Wissen und Erfahrung gelöst werden können. Ebenso erfordern Aufgaben dieses Typs für das Treffen einer Entscheidung mehr als nur basale Fertigkeiten, weil dabei nicht nur Teilhandlungen ausgeführt werden müssen. Neben ICT-spezifischem Wissen erfordert das Treffen von Entscheidungen auch kognitive Fertigkeiten, zum Beispiel um Informationen zu bewerten, Wissen zu transferieren, notwendige Informationen einzuholen und Teilschritte zu planen. Die Items des ICT-Skills-Tests sollen solche alltagstypischen Aufgaben repräsentieren. Die dabei gewonnenen Testwerte sind sollen deshalb als *ICT-spezifischer Fertigkeiten höherer Ordnung* interpretiert werden. Die Testwerteinterpretation ist dann valide, wenn das Treffen der Entscheidungen tatsächlich für die Itemlösung zentral ist, weil beispielsweise die Entscheidungen schwierigkeitsdeterminierend sind. Ebenso sollte das Lösen von ICT-Skills-Items neben dem nötigen Wissen auch kognitive Fertigkeiten erfordern, weil die Entscheidungen Fertigkeiten höherer Ordnung voraussetzen sollten. Auf welchen kognitiven Fertigkeiten ICT-Skills basieren und welche Aufgabencharakteristiken die Schwierigkeit einer Entscheidung beeinflussen und demnach in den Items repräsentiert sein sollen, wird im nächsten Kapitel beschrieben.

2.2 Rahmenkonzeption zur Messung von ICT-Skills (Arbeit 1)

2.2.1 Theoretische Basis für die Itementwicklung

In der psychologischen Leistungsdiagnostik erfolgen Testentwicklungen auf Basis bestimmter psychologischer Theorien, welche es erlauben, Indikatoren für die Itementwicklung abzuleiten. So erfolgt zum Beispiel die Entwicklung von Intelligenztests nach Theorien der fluiden und kristallinen Intelligenz (z.B. Cattell, 1963). Auf deren Basis können anschließend Untertests entwickelt werden, wie zum Beispiel Matrizen-Tests zur Erfassung von schlussfolgerndem Denken als Teil der fluiden Intelligenz (vgl. Wilhelm et al., 2014). Items können auf der Grundlage klar definierter kognitiver Prozesse und daraus abgeleiteter Aufgabencharakteristiken regelbasiert und systematisch konzipiert werden. In Matrizen-Aufgaben können solche Charakteristiken zum Beispiel die Addition oder Subtraktion der Elemente (Hornke & Habon, 1986) und bei geometrischen Analogien die Anzahl der zu verarbeitenden Elemente und durchzufüh-

renden Transformationen sein (Mulholland, Pellegrino & Glaser, 1980). Konstrukte, wie sie in Bildungsvergleichsstudien fokussiert werden, fußen aber meistens weniger auf psychologischen Theorien, sondern eher auf institutionell definierten Wissensdomänen (Watermann & Klieme, 2002, vgl. S.190). Aus diesem Grund sollen im Rahmen dieser Arbeit Theoriebezüge zu fundierten psychologischen Theorien hergestellt werden, um die kognitiven Prozesse bei der Bewältigung von ICT-Aufgaben beschreiben zu können.

Für die Definition des Aufgabenbereichs schlägt Mislevy (2013) im Rahmen seiner evidenzbasierten Erhebungskonzeption vor, zunächst die Domäne zu analysieren und zu organisieren, eine Vorgehensweise, die auch im Rahmen dieser Arbeit verwendet wird. Bisherige Ansätze zur Erfassung von ICT-Skills versuchen den sehr breiten Inhaltsbereich in verschiedene kognitive Aufgaben zu untergliedern, wie zum Beispiel in den Zugriff auf bestimmte Daten oder das Erzeugen von Informationen (International ICT Literacy Panel, 2002). Die Konzeption des International ICT Literacy Panel hat nicht nur andere Konzeptionen inspiriert (National Higher Education ICT Initiative, 2003; International Computer and Information Literacy Study (ICILS), Fraillon et al., 2013), sondern weist auch Überschneidungen zu vielen anderen Konzeptionen auf (Calvani et al., 2009; Eshet-Alkalai, 2004; Ferrari et al., 2012) und kann deshalb als zentral für den Inhaltsbereich ICT angesehen werden. Daher wird diese Unterteilung auch in der vorliegenden Arbeit verwendet und alltägliche ICT-Aufgaben in fünf kognitive Aufgaben unterteilt: in den *Zugriff auf* (*access*), das *Managen* (*manage*), *Integrieren* (*integrate*), das *Bewerten* (*evaluate*) sowie das *Erzeugen* (*create*) von Informationen. Für eine fundierte Testentwicklung sind darüber hinaus Theoriebezüge wichtig (vgl. Borsboom, Mellenbergh & van Heerden, 2004, vgl. S. 1068), um die kognitiven Prozesse, die bei der Bewältigung von ICT-Aufgaben im alltäglichen Kontext ablaufen, für die fünf kognitiven ICT-Aufgaben beschreiben zu können und diese Annahmen in die Itementwicklung einzubeziehen. Eine klassische Kompetenzdomäne in Bildungsvergleichsstudien stellt „Literacy“ dar, definiert als die Fähigkeit, schriftliche Texte zu verwenden, um die eigenen Ziele zu erreichen, Wissen und Potentiale zu verwenden sowie an der Gesellschaft teilzunehmen („understanding, using, reflecting on and engaging with written texts in order to achieve one’s goals, develop one’s knowledge and potential, and participate in society“, OECD, 2011, S.19). Während hier auf Theorien der Leseforschung zurückgegriffen werden kann (z.B. Kintsch, 1998), wird mit ICT-Skills ein Gegenstandsbereich fokussiert, der in erster Linie durch die technische Ent-

wicklung und Präsenz im alltäglichen Leben an Relevanz gewonnen hat und für den es daher keine klassische psychologische Theorie als Basis gibt. Aus diesem Grund besteht das erste konzeptionelle Ziel dieser Arbeit in einer Beschreibung dessen, welche kognitiven Fertigkeiten für das Bearbeiten von ICT-Aufgaben benötigt werden (1), um schwierigkeitsbestimmende Merkmale für die Itementwicklung ableiten zu können (2).

(1) Die zentrale Schwierigkeit in den Items soll das Treffen einer Entscheidung sein (vgl. Abbildung 2). Die gewonnenen Testwerte sollen die Fertigkeit von Personen repräsentieren, solche Entscheidungen zu fällen. Demnach muss zunächst definiert werden, welche kognitiven Fertigkeiten für das Treffen einer derartigen Entscheidung (z.B. für die Weiterleitung einer E-Mail) wichtig sind, um basierend auf den Theorien dieser Fertigkeiten kognitive Schwierigkeiten ableiten zu können. Technische Versiertheit allein genügt nicht, um sich in einer ICT-Umgebung orientieren zu können oder um zu entscheiden, ob eine E-Mail nun weitergeleitet werden soll oder nicht. So muss die E-Mail auch gelesen und verstanden werden. Möglicherweise muss der Benutzer auch noch herausfinden, wie er eine E-Mail innerhalb einer bestimmten Benutzeroberfläche weiterleiten kann. In Einklang mit anderen Rahmenkonzeptionen (Calvani et al., 2009; Fraillon et al., 2013; International ICT Literacy Panel, 2002) wird deshalb davon ausgegangen, dass auch andere kognitive Fertigkeiten benötigt werden, die nicht spezifisch für den ICT-Kontext sind. Neben ICT-spezifischem Wissen werden daher das Verstehen von Text und Grafik sowie das Problemlösen als relevante Fertigkeitsbereiche identifiziert, welche für alle alltäglich anfallenden Aufgaben im ICT-Kontext benötigt werden. Aufgaben, die ICT-spezifische Fertigkeiten höherer Ordnung erfordern, basieren immer auf mindestens einer der beiden generischen Fertigkeiten sowie auf ICT-spezifischem Wissen. Aufgaben, die keine der generischen Fertigkeiten erfordern, sondern allein durch Wissen gelöst werden können, stehen nicht im Fokus dieser Arbeit, weil sie keine Fertigkeiten höherer Ordnung repräsentieren. Hierzu gehören zum Beispiel Routine-Aufgaben, bei denen die Aufgabe und die Benutzeroberfläche bereits vertraut sind, sodass weder Problemlösefertigkeiten noch Fertigkeiten für das Verstehen von Text und Grafik benötigt werden. Aufgaben, für die zwar generische Fertigkeiten, aber kein ICT-spezifisches Wissen nötig sind, wie zum Beispiel reine Leseaufgaben, bilden ebenfalls nicht den Gegenstand der vorliegenden Arbeit, weil sie keine ICT-spezifischen Fertigkeiten erfordern. Im Fokus dieser Untersuchung stehen also Aufgaben, die neben ICT-spezifischem Wissen auch generische Fertigkeiten erfordern.

Nur für solche Aufgaben werden ICT-spezifische Fertigkeiten höherer Ordnung benötigt.

(2) Diese generischen Fertigkeitsbereiche erlauben es nun, auf theoretischen Annahmen fußend schwierigkeitsbestimmende Merkmale bezüglich der Itementwicklung für die fünf kognitiven ICT-Aufgaben abzuleiten. Als theoretische Grundlage für das Problemlösen dient hier die Konzeption von Simon und Newell (1971). Diese Autoren beschreiben, dass das Problemlösen in einem selbst definierten Problemraum abläuft, in dem der Problemlöser unterschiedliche Operatoren wählt, um über verschiedene Knotenpunkte schließlich den Zielzustand zu erreichen. Als Grundlage für das Verstehen von Text und Grafik dienen Modelle des Textverstehens (Kintsch, 1998). Danach werden Buchstaben und Worte verarbeitet, um ein kohärentes Bild des Textinhalts abzubilden, das gemeinsam mit dem Vorwissen zu einem Situationsmodell integriert wird. Beim Integrieren von Informationen aus verschiedenen Dokumenten stellen zum Beispiel die Anzahl der Informationseinheiten sowie das Ausmaß, in dem sich Informationen überlappen und widersprechen, ein Merkmal dar, das die Schwierigkeit eines Integrationsprozesses determiniert (Perfetti, Rouet & Britt, 1999). Diese schwierigkeitsbestimmenden Merkmale aus den generischen Fertigkeitsbereichen werden dann auf ICT-Aufgaben übertragen. Eine typische Aufgabe, die das Integrieren von Informationen erfordert, besteht darin, sich im Internet für einen von zwei Sprachkursen zu entscheiden. Die beschriebenen schwierigkeitsdeterminierenden Aspekte, wie die Anzahl der Informationseinheiten, wären dann die Stellschraube, um Items zu entwickeln, die tatsächlich im Sinne des Integrierens der Informationen schwer sind und die Schwierigkeit der Entscheidung determinieren, etwa bei der Frage: „Welchen Sprachkurs wähle ich?“ Durch solche Stellschrauben soll es möglich sein, Items von unterschiedlicher Schwierigkeit zu erstellen, die jeweils ICT-spezifische Fertigkeiten höherer Ordnung erfordern. Die für die fünf kognitiven ICT-Aufgaben (*Zugriff auf*, das *Managen*, *Integrieren*, *Bewerten* und *Erzeugen* von Informationen) beschriebenen Schwierigkeiten bilden die Grundlage für die Itementwicklung. Weitere schwierigkeitsdeterminierende Merkmale für die anderen kognitiven ICT-Aufgaben werden in der ersten Arbeit detaillierter beschrieben.

Die Bedeutung der Konzeption ist nicht nur für eine theoriegeleitete Itementwicklung, sondern auch für die Validierung der Testwerteinterpretation relevant. So werden die Zusammenhänge mit den generischen Fertigkeiten, Verstehen von Text

und Grafik, Problemlösen und ICT-spezifischem Wissen im Rahmen der Arbeit zur nomothetischen Spanne untersucht (Arbeit 2) und die erarbeiteten schwierigkeitsdeterminierenden Merkmale zur Untersuchung der Konstruktrepräsentation manipuliert (Arbeit 3).

2.2.2 Implementierung verhaltensbasierter Items

Für Tests aus der psychologischen Leistungsdiagnostik, wie in Matrizenaufgaben oder geometrischen Analogien (Hornke & Habon, 1986; Mulholland et al., 1980; Wilhelm et al., 2014), können papierbasierte Stimuli bereits die relevanten kognitiven Prozesse für die Abbildung des Konstruktes hervorrufen. Um dieselben kognitiven Prozesse wie bei der Behandlung von ICT-Aufgaben auch bei der Bearbeitung von Items hervorzurufen, sollten die Aufgaben im besten Falle vergleichbare Reaktionen wie im alltäglichen Leben erlauben. Wie in Abbildung 2 dargestellt wird, werden beim Bearbeiten von ICT-Aufgaben Fertigkeiten verschiedener Ebenen benötigt. Menschen müssen nicht nur dauernd Entscheidungen treffen, sondern immer auch mit ihrer Umgebung interagieren, wenn sie beispielsweise eine E-Mail weiterleiten möchten. Um die kognitiven Prozesse nun realitätsnah abbilden zu können, sollte man auch in der Testsituation mit der Umgebung interagieren. Eine verhaltensbasierte Erhebung mittels Computern ist daher nicht nur naheliegend, sondern geradezu notwendig. Solche innovativen Erhebungsmethoden sollten die Repräsentation des Konstrukts in den Testwerten erhöhen (Sireci & Zenisky, 2006, S. 329). Mislevy (2013) hebt das Problem der Konstruktion einer solchen Simulationsumgebung hervor. Die Schwierigkeit liegt hier darin, zu entscheiden, welche Aspekte der Realität nun für eine adäquate Repräsentation des Konstrukts abgebildet werden müssen und welche weggelassen werden können, etwa weil diese die Erstellung der Umgebung in erster Linie verkomplizieren können. Es gilt also, einen Mittelweg zwischen Authentizität und Machbarkeit im Sinne zeitlicher und finanzieller Ressourcen zu finden. Für das Problem der Umsetzung soll immer die leitende Frage gestellt werden, ob die kognitiven Prozesse angemessen hervorgerufen werden. Denn nur wenn die Simulationsumgebung auf die Testpersonen authentisch wirkt, kann davon ausgegangen werden, dass die kognitiven Prozesse ähnlich wie in tatsächlich anfallenden Aufgaben ablaufen. Wenn es eine Aufgabe erfordert, mehrere E-Mails hinsichtlich ihrer Glaubwürdigkeit zu beurteilen, sollten Personen tatsächlich alle E-Mails gesehen und beurteilt haben, da nur so eine Interpretation des Testwertes im Sinne von Glaubwürdigkeitsbeurteilung valide ist.

Um dies zu gewährleisten, sollen die Aufgaben so authentisch wie möglich umgesetzt werden. Solange sich Personen wie intendiert verhalten, sollten die Limitationen der Simulationsumgebung nicht zu sehen sein (z.B. führt eine Suche im Browser nur dann zu Ergebnissen, wenn sie tatsächlich zur Aufgabenlösung beiträgt). Dies erfordert auch, dass sowohl der richtige als auch der falsche Lösungsweg vollständig implementiert sind. Instruktionen sollen hierbei darauf hinleiten, dass sich Personen wie intendiert verhalten und die intendierten kognitiven Prozesse auch tatsächlich durchgeführt werden. Die Authentizität soll dadurch unterstützt werden, dass die Antwortabgabe immer innerhalb der Aufgabe erfolgt, zum Beispiel durch das tatsächliche Weiterleiten einer E-Mail. Eine möglichst authentische Simulation in Aufgabe und Antwortabgabe soll dazu führen, dass die kognitiven Prozesse – zum Beispiel beim Bewerten von Informationen – realitätsnah ablaufen. Eine sehr realitätsnahe Gestaltung einer Aufgabe führt aber auch dazu, dass sich die resultierenden Items in ihrer Oberfläche stark unterscheiden, zum Beispiel wenn E-Mail-Postfächer, Websites und Textverarbeitungsprogramme simuliert werden.

2.3 Heterogene Items für die Erfassung von ICT-Skills

Aus den formulierten Rationalen für die Itementwicklung und -implementierung resultieren sehr heterogene Items. Zurückführen lässt sich dies zum einen auf die Breite des Gegenstandsbereichs ICT-Skills, wie sie in dieser Studie erfasst werden sollen. Jede Informationsaufgabe, vom Zugreifen auf Informationen bis zum Erzeugen von Informationen (vgl. International ICT Literacy Panel, 2002), kann mithilfe von Technologien durchgeführt werden und die entwickelten Items sollen alle fünf kognitiven ICT-Aufgaben abbilden. Zum anderen soll dieses sehr breite Spektrum an Aufgaben gleichzeitig verhaltensbasiert untersucht werden. Dies bedeutet, dass nicht nur der Inhalt (z.B. *Zugriff auf* vs. *Erzeugen* von Informationen), sondern auch das Format (z.B. Browser vs. Tabellenkalkulationssoftware) der resultierenden Items sehr unterschiedlich ist. Es kommt hinzu, dass eine realitätsnahe Umgebung mit authentischen Reaktionen dazu führt, dass die Reaktionen in den verschiedenen Items sehr unterschiedlich sind. So mögen es einige Items erfordern, eine E-Mail zu schreiben. Dieses Verhalten wäre als ein eher konstruktives Antwortformat zu verstehen. Andere Items könnten es hingegen nahelegen, für eine ausgewählte Website ein Lesezeichen zu setzen, was als eher selektives Antwortformat anzusehen wäre (Scalise & Gifford, 2006). Jedes Item mag

somit unterschiedliche kognitive Prozesse zum Treffen der Entscheidung und zur Antwortabgabe erfordern.

Die Heterogenität der Items kann auch durch das Modell in Abbildung 2 verdeutlicht werden. Würden alle entwickelten Items durch ein Handlungsmodell beschrieben, dann ergäbe sich eine Vielzahl von unterschiedlichen Modellen, möglicherweise ein eigenes Modell pro Item, während in einem homogenen Item-Set mehrere Items durch ein gemeinsames Handlungsmodell beschrieben werden könnten. Jedes Item würde in einem heterogenen Itemset demnach einen eigenen Itemtyp darstellen, während Items in homogenen Item-Sets einem oder wenigen Itemtypen zuzuordnen wären (siehe Abbildung 3). Ein Beispiel für homogene Items im ICT-Bereich wäre zum Beispiel das Bewerten von Online-Informationen („Evaluating Online Information"; Pfaff & Goldhammer, 2011, September; siehe auch Hahnel, Goldhammer, Naumann & Kröhne, 2016), durch das mehrere Items einem Itemtyp zugeordnet werden können. Alle Items eines Itemtyps sind beispielsweise auf einer Suchergebnisseite angeordnet, aus der jeweils ein Link ausgewählt werden muss. Die Reaktion wäre in allen Items dieselbe. Die einzelnen Items unterschieden sich dann möglicherweise nur in der Anzahl der zu vergleichenden Links oder aber der Güte der Distraktoren zum richtigen Link. Für ICT-Skills, wie sie in dieser Arbeit untersucht werden sollen und auch in anderen Studien erforscht wurden (vgl. International ICT Literacy Panel, 2002), würde nur ein Item von diesem Itemtyp in einem Test enthalten sein. Denn ICT-Skills erfordern nicht nur das Bewerten, sondern auch das Zugreifen auf, Managen, Integrieren und Erzeugen von Informationen. Darüber hinaus soll das Bewerten von Informationen auch durch eine Vielzahl von Indikatoren – und nicht nur durch das Bewerten von Suchergebnisseiten im Browser – abgebildet werden, zum Beispiel auch von E-Mails.

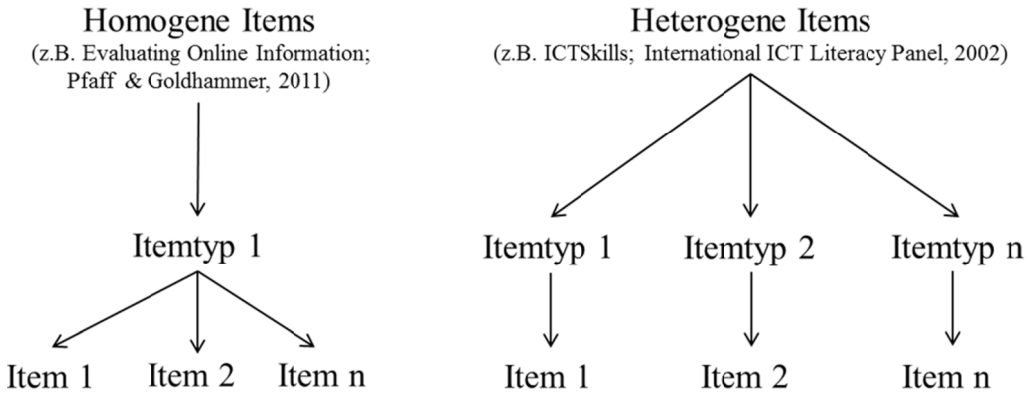


Abbildung 3. Homogene und heterogene Items.

2.4 Konzept zur Konstruktvalidierung

In der vorliegenden Arbeit orientiert sich das Verständnis von Validität an den herausgegebenen Standards der amerikanischen Fachverbände AERA, APA und NCME (AERA, APA & NCME, 2014). Diese verstehen Validität nicht als Eigenschaft eines Tests, sondern beziehen sie auf eine bestimmte Interpretation der Testwerte, je nachdem, wofür diese Werte genutzt werden sollen. Die Untersuchung der Validität stellt einen Prozess dar (Kane, 2013), bei dem empirische Ergebnisse die Evidenzen für eine valide Interpretation anreichern, wobei jede neue Interpretation und weitere Verwendung der Testwerte eine erneute Sammlung an Evidenzen erfordert. Die Wahl der zu betrachtenden Evidenzquellen orientiert sich an der vorab definierten Testwerteinterpretation, also einer Aussage darüber, wie Testergebnisse interpretiert werden sollen. Das Validierungskonzept hat also zum Ziel, die Validität der Testwerteinterpretation zu untersuchen. Diese Interpretation wird gestützt, wenn die Testwerte – wie in der Rahmenkonzeption postuliert (vgl. Abschnitt 2.2.1) – mit den beschriebenen Konstrukten zusammenhängen und die abgeleiteten schwierigkeitsbestimmenden Merkmale die Schwierigkeit determinieren. Diese beiden Annahmen können mit den zwei von Embretson (1983) beschriebenen Verfahren, dem Ansatz der nomothetischen Spanne und der Konstruktrepräsentation, untersucht werden. Für die Verwendung heterogener Items ist es jedoch notwendig, klassische Validierungsverfahren anzupassen und zu erweitern. Daher werden im Folgenden zunächst zwei Validierungsansätze beschrieben (2.4.1) (Embretson, 1983), um dann auf die jeweiligen Anpassungen und Weiterentwicklungen für heterogene Items einzugehen (2.4.2 und 2.4.3).

2.4.1 Klassische Verfahren der Validierung: Nomothetische Spanne und Konstruktrepräsentation

Embretson (1983) beschreibt zwei Ansätze zur Überprüfung der Validität. Während sich der Ansatz der Konstruktrepräsentation auf Unterschiede zwischen Items konzentriert, bezieht sich der Ansatz der nomothetischen Spanne auf Unterschiede zwischen Personen. Die Untersuchung der nomothetischen Spanne stellt ein klassisches Verfahren dar, bei der ein Messverfahren in ein breiteres nomologisches Netzwerk eingeordnet wird (Cronbach & Meehl, 1955). Auch wenn heute nicht mehr von der Validität eines Tests, sondern der Validität der Testwerteinterpretation für einen bestimmten Zweck gesprochen wird, bildet der Zusammenhang der Testwerte mit anderen Variablen auch in den aktuellen Standards für Validität eine der Evidenzquellen (vgl. *evidence*

based on relations to other variables; AERA, APA & NCME, 2014). Bei der Untersuchung der nomothetischen Spanne werden Testwerte des zu messenden Konstrukts mit anderen Maßen ins Verhältnis gesetzt. Auf diese Weise kann untersucht werden, ob die Testwerte wie theoretisch angenommen mit anderen Maßen korrelieren (konvergente Evidenz) oder ob dies nicht der Fall ist (diskriminante Evidenz). Es können Aussagen darüber getroffen werden, inwiefern Konstrukte voneinander abgrenzbar sind, beispielsweise komplexes Problemlösen von Intelligenz (Kretzschmar, Neubert, Wüstenberg & Greiff, 2016), oder inwiefern Testwerte neu entwickelter Tests, die zum Beispiel verschiedene Teilfähigkeiten des Leseverstehens abbilden sollen, mit anderen Instrumenten korrelieren (Richter, Isberner, Naumann & Kutzner, 2012).

Für die Überprüfung der Konstruktrepräsentation könnte nach Embretson (1983) untersucht werden, inwiefern schwierigkeitsdeterminierende Aufgabenmerkmale, die bestimmte kognitive Prozesse hervorrufen sollen, mit der resultierenden Itemschwierigkeit zusammenhängen. Hierbei werden schwierigkeitsdeterminierende Aufgabencharakteristiken verwendet, die über Items vergleichbar oder sogar quantifizierbar sind. In den aktuellen Standards für Validität werden solche Evidenzen als auf Antwortprozessen basierend beschrieben (*evidence based on response processes;* AERA, APA & NCME, 2014), weil Annahmen über kognitive Prozesse in den Aufgaben empirisch überprüft werden. Schwierigkeitsbestimmende Merkmale können bei Matrizen-Aufgaben zum Beispiel die Addition oder Subtraktion der Elemente sein (Hornke & Habon, 1986). Bei geometrischen Analogien wäre dies die Anzahl der zu verarbeitenden Elemente und durchzuführenden Transformationen (Mulholland et al., 1980). Dieser Ansatz wurde bereits in verschiedenen Studien mit homogenem Itemmaterial angewendet, zum Beispiel für mentale Rotationsaufgaben (Caissie, Vigneau & Bors, 2009), Problemlöseaufgaben (Greiff, Krkovic & Nagy, 2014; Stadler, Niepel & Greiff, 2016) oder Leseverständnisaufgaben (Hartig & Frey, 2012).

2.4.2 Untersuchung der nomothetischen Spanne bei heterogenen Items (Arbeit 2)

Basierend auf der Testwerteinterpretation sollen die Testwerte ICT-Skills höherer Ordnung abbilden. Für das Lösen von Aufgaben, die solche Fertigkeiten erfordern, werden in der Rahmenkonzeption ICT-spezifisches Wissen und generische Fertigkeiten, Verstehen von Text und Grafik sowie Problemlösen als wichtige Komponenten spezifiziert. Für die Erhebung von ICT-Skills werden heterogene Items verwendet und jedes

Item stellt potentiell einen eigenen Itemtyp dar (vgl. Abbildung 3). Items können zum Beispiel das Managen von Informationen in Tabellenkalkulationssoftware oder das Bewerten einer Website hinsichtlich ihrer Glaubwürdigkeit erfordern. Notwendige Fertigkeiten, etwa Leseverstehen oder Wissen, können demnach je nach Item sehr unterschiedlich sein, etwa weil die Textmenge variiert oder weil je nach Item unterschiedliche Wissensbestände für die Itemlösung relevant sind. Bei der Untersuchung der nomothetischen Spanne werden Testwerte miteinander in einen Zusammenhang gestellt und Zusammenhänge zwischen allen Items betrachtet. Bei heterogenen Items kann aus den obigen Gründen aber nicht automatisch davon ausgegangen werden, dass der Zusammenhang mit anderen Testwerten für jedes der realisierten Items derselbe ist. Auch wenn der Zusammenhang der Testwerte mit anderen Konstrukten gemäß der Erwartung ausfällt, bedeutet dies nicht automatisch, dass es auch für jedes der einzelnen Items gilt. Wenn die Testwerte wie intendiert mit anderen Variablen zusammenhängen, kann dies also bei sehr heterogenen Items daran liegen, dass einzelne Items besonders hoch und andere niedrig mit diesen Variablen zusammenhängen.

Dies ist dann problematisch, wenn diese Variablen als konstrukt-bestimmend verstanden werden. In dieser Arbeit sollen Testwerte als ICT-spezifische Fertigkeiten höherer Ordnung verstanden werden, was bedeutet, dass jedes Item sowohl ICT-spezifisches Wissen, aber darüber hinaus auch generische Fertigkeiten erfordern sollte, um die Testwerte wie intendiert interpretieren zu können. Würden einige Items in erster Linie nur Leseverstehen oder Problemlösen und andere in erster Linie nur ICT-spezifisches Wissen erfordern, würde die Testwerteinterpretation trotz eines Zusammenhangs mit allen drei Variablen über alle Items hinweg nicht gestützt.

Inwiefern die Heterogenität der Items tatsächlich zu unterschiedlichen Zusammenhängen für verschiedene Items führt, kann darüber hinaus durch Itemmerkmale erklärt werden. Ausgehend von der Annahme, dass das Leseverstehen aufgrund benötigter Leseprozesse in den Items mit der Lösung der ICT-Skills-Items zusammenhängt, könnte die Textmenge zur Erklärung dieses Zusammenhangs dienen. Eine solche Erklärung würde korrelative Befunde durch erklärende Anteile untermauern, gerade wenn unterschiedliche kognitive Prozesse in den Aufgaben die Konstruktinterpretation nicht infrage stellen, weil sie eben eine Eigenschaft des Gegenstandsbereiches sind.

2.4.3 Untersuchung der Konstruktrepräsentation bei heterogenen Items (Arbeit 3)

Um Testwerte als ICT-Skills höherer Ordnung interpretieren zu können, sollten nach dem in Abbildung 2 beschriebenen Modell die zu treffenden Entscheidungen die Schwierigkeit in Items determinieren und für die intendierte Testwerteinterpretation verantwortlich sein. Je schwerer eine solche Entscheidung ist, desto schwerer sollte ein Item sein, und Items, die keine Entscheidungen beinhalten, sollten das intendierte Konstrukt nicht mehr repräsentieren. Um dies näher zu bestimmen, würde der Ansatz der Konstruktrepräsentation nach Embretson nun erfordern, Merkmale, welche die Schwierigkeit einer solchen Entscheidung beschreiben, in allen Items zu quantifizieren und zu untersuchen, ob diese die Itemschwierigkeit erklären können.

Ein solches Merkmal könnte in den Items zur Bewertung von Online-Informationen die Diskrepanz der Glaubwürdigkeit von den Distraktoren zum richtigen Link sein, wobei eine größere Diskrepanz zu leichteren Items führen sollte (vgl. Abschnitt 2.3 „Evaluating Online Information“; Pfaff & Goldhammer, 2011, September; siehe auch Hahnel et al. 2016). In dem hier entwickelten Test würde es aber nur ein solches Item geben und andere Items würden nicht nur eine Glaubwürdigkeitsbewertung erfordern (Bewerten von Informationen), sondern auch andere kognitive ICT-Aufgaben wie zum Beispiel die Auswahl eines Suchbegriffs für eine Internetsuche (Zugriff auf Informationen) oder die sinnvolle Benennung eines Dokuments (Erzeugen von Informationen). Itemmerkmale sind zwischen Items in heterogenen Itemsets also nicht vergleichbar. Deshalb sollen die Merkmale nicht zwischen Items verglichen, sondern manipuliert werden. Dadurch werden für die Items neue Item-Varianten erstellt, die sich nur in dem manipulierten Itemmerkmal unterscheiden. In homogenen Itemsets hingegen würde jedes Item bereits eine Item-Variante des anderen Items darstellen. Analog zu Embretsons Ansatz (1983) soll untersucht werden, ob sich durch die Manipulation des Merkmals die Itemschwierigkeit ändert. Zudem soll der Frage nachgegangen werden, ob sich durch die Manipulation des Merkmals auch das gemessene Konstrukt verändert. Die in der ersten Arbeit abgeleiteten schwierigkeitsdeterminierenden Merkmale erlauben es, Annahmen über kognitive Prozesse abzuleiten, die zur Lösung der Items benötigt werden (vgl. Kane, 2013, S. 38). Der für diese Arbeit verwendete Ansatz basiert auf dem Informationsverarbeitungsansatz und stellt Annahmen über kognitive Prozesse mit der Itembearbeitung in einen Zusammenhang (vgl. Borsboom et al., 2004). In den experimentellen Manipulationen wurden eben die genannten Annahmen über die kognitiven

Prozesse genutzt, um diese zu verändern (siehe *Change*-Ansatz) oder Teile daraus zu eliminieren (siehe *Eliminate*-Ansatz). Diese Manipulationen setzen an den zu treffenden Entscheidungen an. Um höhere Testwerte als ICT-spezifische Skills höherer Ordnung interpretieren zu können, sollte die Herausforderung, in einem ICT-Skills-Item tatsächlich durch die Schwierigkeit, Entscheidungen richtig zu treffen, hervorgerufen werden, und nicht nur durch die einzelnen durchzuführenden Lösungsschritte. Ein Item sollte also umso schwerer sein, je schwerer die zu treffende Entscheidung ist.

(1) Der *Change*-Ansatz ist zur Untersuchung der Frage geeignet, inwiefern ICT-spezifische Fertigkeiten in den Testwerten repräsentiert sind. Hierbei sollen die Aufgabencharakteristiken, welche die ICT-spezifische Schwierigkeit bestimmen sollen (z.B. die Bewertung einer E-Mail auf Basis von Spam-Indikatoren) und in Arbeit 1 herausgearbeitet wurden, schwerer oder leichter gemacht werden. Nach Abbildung 2 würde dies die Veränderung einer Entscheidung bedeuten. Führt eine solche Manipulation zu einer veränderten Itemschwierigkeit, determiniert die zu treffende Entscheidung tatsächlich die Schwierigkeit eines Items. Eine solche Manipulation im Sinne des Konstrukts sollte ebenso nicht zu einer veränderten Testwerteinterpretation führen, also Testwerte sollten nach wie vor ICT-spezifische Fertigkeiten höherer Ordnung repräsentieren.

(2) Der *Eliminate*-Ansatz eignet sich zur Untersuchung der Frage, inwiefern Fertigkeiten höherer Ordnung tatsächlich in den Testwerten repräsentiert sind. Für eine solche Prüfung soll die zu interessierende Fertigkeitsebene aus den Items eliminiert werden. In Abbildung 2 würde dies alle Entscheidungen betreffen, die in den Items enthalten sind. Solche Eliminate-Items sollten leichter sein, weil die schwierigkeitsdeterminierenden Merkmale nicht mehr vorhanden sind. Darüber hinaus sollten solche Items nicht mehr dasselbe Konstrukt – ICT-spezifische Fertigkeiten höherer Ordnung – repräsentieren, weil die Merkmale, die eben Fertigkeiten höherer Ordnung erfordern sollten, nicht mehr in den Items enthalten sind.

3 FORSCHUNGSFRAGEN

Ausgehend von den konzeptionellen Arbeiten werden im Folgenden die Forschungsfragen und Hypothesen formuliert. Das Ziel der empirischen Analysen der ersten Arbeit besteht darin, zu überprüfen, inwiefern die zur Itementwicklung und -implementierung entwickelte Rahmenkonzeption zur Erstellung der Items geeignet war. Das Ziel der zweiten und dritten Arbeit ist es, dem Validierungskonzept folgend Evidenzen zu sammeln für die Validität der Testwerteinterpretation, Testwerte als ICT-spezifische Fertigkeiten höherer Ordnung zu interpretieren. Zur besseren Übersicht sind in Tabelle 1 die konzeptionellen Beiträge, welche in Kapitel 2 beschrieben wurden, gemeinsam mit den empirischen Beiträgen den drei eigenständigen Arbeiten zugeordnet.

Tabelle 1. Zuordnung der konzeptionellen und empirischen Beiträge zu den drei Arbeiten. Die Bezüge zu den Abschnitten beschreiben, an welcher Stelle im Rahmentext die Beiträge der drei Arbeiten verortet sind.

	Arbeit	Konzeptionelle Beiträge		Empirische Beiträge	
1	Rahmenkonzeption zur verhaltensbasierten Erhebung von ICT-Skills *[A framework for the performance-based testing of ICT skills]*	Theoretische Basis für die Itementwicklung	2.2.1	Forschungsfrage 1a	4.2.1
		Rationale für die Itemimplementierung verhaltensbasierter Items	2.2.2	Forschungsfrage 1b	
2	Konvergente Evidenz für die Validität eines verhaltensbasierten ICT-Skills-Tests *[Convergent evidence for validity of a performance-based ICT skills test]*	Aufgabenzentrierter Ansatz zur Überprüfung der nomothetischen Spanne bei heterogenen Itemsets	2.4.2	Hypothesen 2a-c	4.2.2
3	Experimentelle Validierungsstrategien für heterogene und computerbasiert erhobene Items *[Experimental validation strategies for heterogeneous computer-based assessment items]*	Beschreibung zweier Ansätze zur Überprüfung der Konstruktrepräsentation bei heterogenen Itemsets - Change - Eliminate	2.4.3	Hypothesen 3a,b 3c,d	4.2.3

Bemerkung: Die Benennung der Forschungsfragen und Hypothesen in diesem Rahmentext weicht zur besseren Verständlichkeit von der in den drei eigenständigen Arbeiten verwendeten Benennung ab.

3.1 Erprobung der Rahmenkonzeption (Arbeit 1)

Obgleich der Schwerpunkt der ersten Arbeit auf der konzeptionellen Entwicklung der Rahmenkonzeption liegt, sollen darüber hinaus empirische Evidenzen für die Anwendbarkeit der Rahmenkonzeption gewonnen werden. Im Folgenden werden zwei Forschungsfragen formuliert, um zu untersuchen, inwieweit die in Abschnitt 2.2 entwickelte Rahmenkonzeption zur Itementwicklung und -implementierung geeignet ist. Ein zentraler Aspekt der Rahmenkonzeption stellt die Herleitung der schwierigkeitsdeterminierenden Merkmale für die fünf kognitiven ICT-Aufgaben dar (z.B. den Zugriff auf Informationen), welche als Stellschrauben für die jeweilige Itemschwierigkeit in der Itementwicklung dienen sollten. Als Konsequenz sollte es möglich sein, durch die Nutzung dieser Merkmale systematisch Items mit vergleichbaren Schwierigkeiten für alle fünf kognitiven Aufgaben *(Zugriff auf*, das *Managen*, *Integrieren*, *Bewerten* und *Erzeugen von Informationen*) zu entwickeln, und zwar sowohl im niedrigeren als auch im höheren Schwierigkeitsbereich.

Forschungsfrage 1a: Decken die entwickelten Items einen vergleichbaren Bereich an Itemschwierigkeiten für die fünf kognitiven ICT-Aufgaben ab?

Die simulationsbasierten Items sollten dieselben kognitiven Prozesse wie in alltagstypischen Aufgaben hervorrufen. Deshalb wurden Rationale für die Implementierung entwickelt, zum Beispiel hinsichtlich einer authentischen Umgebung oder einer realistischen Art der Antwortabgabe (vgl. Abschnitt 2.2.2). Das Verhalten in den Items kann ein Indiz dafür sein, ob die kognitiven Prozesse wie intendiert abgelaufen sind, was beispielsweise dann der Fall ist, wenn für eine Entscheidung zwischen zwei Sprachschulen im Internet tatsächlich beide Websites betrachtet wurden. Ist dies nicht erfolgt, kann die Anzahl der verwendeten Lösungsschritte ein Indiz hierfür sein, wenn zum Beispiel weniger Interaktionen verwendet wurden als für die Betrachtung beider Websites nötig gewesen wären. Selbst wenn das Item richtig gelöst werden konnte, würde dies darauf hindeuten, dass es nicht richtig funktioniert, weil die Entscheidung für eine Website nicht auf Grundlage beider Websites getroffen wurde. Daher soll die Anzahl der verwendeten Interaktionen – nach Abbildung 2 die Lösungsschritte – Aufschluss darüber geben, ob die kognitiven Prozesse wie intendiert durchgeführt wurden.

Forschungsfrage 1b: Verhält sich der Großteil der Testteilnehmer/innen in den simulierten Aufgaben wie intendiert?

3.2 Untersuchung der nomothetischen Spanne (Arbeit 2)

In dieser Arbeit sollen Testwerte als ICT-spezifische Fertigkeiten höherer Ordnung interpretiert werden. Die hierfür erforderlichen Fertigkeiten für das Lösen solcher Aufgaben wurden in der Rahmenkonzeption als ICT-spezifisches Wissen definiert, gemeinsam mit den generischen Fertigkeiten des Problemlösens und des Verstehens von Text und Grafik. Höhere Testwerte sollten demnach mit höheren Werten in diesen drei Variablen einhergehen, damit die Testwerte wie intendiert interpretiert werden können.

Hypothese 2a: Unabhängige positive Effekte von ICT-spezifischem Wissen, Problemlösen sowie Verstehen von Text und Grafik auf die Wahrscheinlichkeit, ICT-Skills-Items zu lösen, stützen die intendierte Konstruktinterpretation.

Aufgrund der heterogenen Items soll das postulierte Zusammenspiel der drei Fertigkeiten auch auf der Itemebene betrachtet werden. Items, welche kein ICT-spezifisches Wissen benötigen, repräsentieren auch keine ICT-spezifischen Fertigkeiten. Und Items, die weder Problemlösefertigkeiten noch Verstehen von Text und Grafik erfordern, stellen reine Routine-Aufgaben dar und erfordern keine Fertigkeiten höherer Ordnung.

Hypothese 2b: Positive Zusammenhänge für die Wahrscheinlichkeit, einzelne ICT-Skills-Items zu lösen, mit ICT-spezifischem Wissen und eine oder beide der generischen Variablen (Problemlösen und Verstehen von Text und Grafik) stützen die intendierte Konstruktinterpretation.

Der Zusammenhang von ICT-Skills mit dem Verstehen von Text und Grafik sowie mit dem Problemlösen soll auf ähnlichen kognitiven Prozessen beim Lösen beider Aufgaben basieren. Indikatoren für Problemlöse- und Leseprozesse können als Erklärungsvariablen für die Zusammenhänge mit den ICT-Skills-Items dienen. Eine größere Textmenge in den Items sollte zum Beispiel zu einem höheren Zusammenhang mit dem Leseverstehen führen, da solche Items mehr Leseverstehensprozesse erfordern sollten. Analog hierzu sollte eine hohe intrinsische Komplexität mehr Problemlöseprozesse erfordern und solche Items sollten stärker mit Problemlösefertigkeiten zusammenhängen.

Hypothese 2c: Die Konstruktinterpretation wird weiterhin gestützt, wenn Itemmerkmale die Zusammenhänge der generischen Variablen (Problemlösen sowie Verstehen von Text und Grafik) mit der Wahrscheinlichkeit, ICT-Skills-Items zu lösen, erklären können.

3.3 Untersuchung der Konstruktrepräsentation (Arbeit 3)

Zur Untersuchung der Konstruktrepräsentation (Embretson, 1983) wurden zwei experimentelle Ansätze entwickelt, die es erlauben, diesen Ansatz auf heterogene Items zu übertragen. Da die Entscheidungen in den Items ICT-spezifische Fertigkeiten höherer Ordnung erfordern (vgl. Abbildung 2) und die Stellschraube für die Itemschwierigkeit bilden sollten, stellen diese die zu manipulierenden Itemmerkmale dar. Hierbei werden mit dem *Change*-Ansatz die zu treffenden Entscheidungen verändert, und zwar entweder erleichtert oder erschwert, um zu untersuchen, ob die Testwerte ICT-spezifische Fertigkeiten repräsentieren. So veränderte Items sollen zudem dasselbe Konstrukt wie die originalen Items messen, was sich durch einen unveränderten Zusammenhang mit konstrukt-verwandten Personenvariablen untermauern lässt.

Hypothese 3a: Eine Veränderung der Aufgabenmerkmale (Change-Manipulation) führt zu einer Veränderung der Itemschwierigkeiten in die intendierte Richtung.

Hypothese 3b: Die Veränderung der Aufgabenmerkmale (Change-Manipulation) führt zu einem unveränderten Zusammenhang mit Personenvariablen.

Mit dem *Eliminate*-Ansatz werden alle Entscheidungen in einem Item eliminiert, um zu untersuchen, ob die Testwerte Fertigkeiten höherer Ordnung repräsentieren. Die Entscheidungen werden als zentral für die Konstruktrepräsentation angesehen, da sie Fertigkeiten höherer Ordnung erfordern sollten, die über basale Computerfertigkeiten hinausgehen. Items, die es nicht erfordern, Entscheidungen auf Grundlage ICT-spezifischer Fertigkeiten höherer Ordnung zu treffen, sollen demnach nicht nur leichter sein, sondern vor allem auch ein anderes Konstrukt messen.

Hypothese 3c: Eliminate-Items sind leichter als die originalen Items.

Hypothese 3d: Die Veränderung der Aufgabenmerkmale (Eliminate-Manipulation) führt zu einem veränderten Zusammenhang mit Personenvariablen.

4 EMPIRISCHE ANALYSEN

Die Analysen aller drei hier vorgelegten Arbeiten basieren auf einer gemeinsamen Datenerhebung, die innerhalb des vom BMBF geförderten Verbundprojekts CavE-ICT-PISA[1] im Rahmen der Initiative zur Förderung von Forschungsvorhaben in Ankopplung an Large-Scale-Assessments (LSA) durchgeführt wurde. Im Folgenden werden zunächst methodische Aspekte erläutert (4.1) und dann die empirischen Ergebnisse nacheinander dargestellt (4.2).

4.1 Methode

4.1.1 Itempool

Im Zentrum der Studien steht der ICT-Skills-Test, welcher auf Basis der Rahmenkonzeption (vgl. Arbeit 1) im Projekt CavE-ICT-PISA entwickelt wurde (Wenzel et al., 2016). Der gesamte Itempool bestand aus 70 Items und wurde von sieben Itementwicklern erstellt, welche die Itementwürfe gegenseitig hinsichtlich ihrer inhaltlichen Passung und technischen Umsetzbarkeit kommentiert und optimiert haben. Die Items wurden in der Simulationsumgebung mit dem CBA ItemBuilder implementiert (Rölke, 2012), inklusive einer direkt nach Antwortabgabe ablaufenden dichotomen Bewertung. Die Items wurden so entwickelt, dass diese eine der fünf kognitiven ICT-Aufgaben repräsentieren (den *Zugriff auf*, das *Managen*, *Integrieren*, *Bewerten* sowie das *Erzeugen von Informationen*). Darüber hinaus wurden verschiedene Applikationen umgesetzt, zum Beispiel Web-Browser, E-Mail-Postfach, Dateimanager sowie Textverarbeitungs- oder Tabellenkalkulationssoftware. Itembeispiele werden in der ersten Arbeit beschrieben. Ebenso wird ein Beispiel eines implementierten Items in der dritten Arbeit dargestellt.

Zur Untersuchung der Konstruktrepräsentation in der dritten Arbeit wurden von den 70 ICT-Skills-Items 40 Items nach dem Change-Ansatz und 20 Items nach dem Eliminate-Ansatz verändert. Für die Change-Varianten wurden die zu treffenden Entscheidungen für 30 Items erleichtert und für 10 Items erschwert. Hierfür wurde zum Beispiel der Verfasser einer zu bewertenden E-Mail in einen weniger vertrauenswürdi-

[1] CavE-ICT-PISA steht für Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologiebezogener Fähigkeiten (ICT-Skills) in PISA. Der Projektbericht kann eingesehen werden unter: http://edok01.tib.uni-hannover.de/edoks/e01fb15/838771823.pdf (letzter Aufruf: 31.8.2017)

gen Verfasser geändert, was die Bewertung dieser E-Mail erleichtern sollte. Für die Eliminate-Varianten wurden die Lösungen der Entscheidungen bereits in der Instruktion gegeben, zum Beispiel welche E-Mail weitergeleitet oder welche Website zu den Lesezeichen hinzuzufügen ist. So sollten die Items keine Fertigkeiten höherer Ordnung mehr erfordern, da nur noch Teilhandlungen und einzelne Lösungsschritte durchgeführt werden müssen (vgl. Abbildung 2).

4.1.2 Datenerhebung und Prozeduren

Die Datenerhebung fand im Juni und Juli 2014 an ausgewählten Schulen in Baden-Württemberg und Rheinland-Pfalz statt. Diese Schulen wurden zufällig vom IEA Data Processing Center (DPC) ausgewählt. Die technische Ausstattung der jeweiligen Einrichtung und die Schulleitung bildeten das Kriterium dafür, welche Schule an der Erhebung teilnehmen sollte. Zudem entschieden die Schüler/innen sowie die Eltern über die Teilnahme Einzelner. Die gesamte Stichprobe umfasste 983 Schüler/innen, die im Durchschnitt 15 Jahre alt (M = 15,29, SD = 0,66) und zur Hälfte männlich waren (51% männlich, 46% weiblich, 3% nicht spezifiziert). Von den 34 Schulen waren elf Schulen Gymnasien und die übrigen Haupt-, Gesamt-, Real- sowie Realschulen plus. Die Testung bestand aus zwei Teilen zu je 60 Minuten und einer dazwischen liegenden zehnminütigen Pause. Zu Beginn bearbeiteten alle Schüler/innen Übungsaufgaben, um sich mit der Simulationsumgebung vertraut zu machen. Die Zuweisung zu den verschiedenen Testheften erfolgte randomisiert (vgl. Tabelle 2). Die ICT-Skills-Items wurden in elf verschiedenen Testheften in einem unvollständig balancierten Testheftdesign administriert (vgl. Wenzel et al., 2016). Die Schüler/innen bearbeiteten bis zu 33 Items.

Tabelle 2. Zuordnung der in dieser Arbeit verwendeten Variablen zu den Testbedingungen.

	Teil 1 (60 min)	Teil 2 (60 min)	N = 983
1		ICT-Nutzung	284
2	ÜBUNG ICT-Skills-Items	PAUSE ICT-Nutzung, ICT-spezifisches Wissen, Eliminate-Items	220
3	ÜBUNG ICT-Skills-Items	PAUSE ICT-Nutzung, ICT-spezifisches Wissen, Leseverstehen, Problemlösen	269
4	ÜBUNG Change-Items	PAUSE ICT-Nutzung, ICT-spezifisches Wissen	210

Bemerkung: Aus den vier Gruppen der ersten Spalte ergeben sich die Stichprobenzahlen für die drei Arbeiten. In der Arbeit 1 wurden alle ICT-Skills-Items untersucht. Daher waren Personen der Gruppen 1-3 Teil der Analysen. Die Analysen der Arbeit 2 basierten nur auf Personen der Gruppe 3, die alle relevanten Instrumente bearbeitet hatten. Die Arbeit 3 basiert auf der gesamten Stichprobe, da auch die Change-Items Teil der Analysen waren.

4.1.3 Variablen

In den Tabellen 3-5 werden die verwendeten Variablen tabellarisch in knapper Form aufgeführt. Dabei werden sie nach Personen-, Item- und Prozessmerkmalen sortiert.

Tabelle 3. Verwendete Personenmerkmale für die empirischen Analysen.

Konstrukt	Instrument	Itemformate	Variable basiert auf	Eigenschaften
ICT-spezifisches Wissen[2,3]	TECOWI (Test zur Erfassung des theoretischen Computerwissens; Richter, Naumann & Horz, 2010)	20 Multiple-Choice-Fragen	Summenwert	Cronbachs alpha: $\alpha = 0{,}68$ $M = 7{,}5$ ($SD = 3{,}28$)
Verstehen von Text und Grafik[2]	Lesegeschwindigkeits- und Verständnistests (LGVT; Schneider, Schlagmüller & Ennemoser, 2007)	23 Lücken mit je 3 Worten zur Auswahl	Summenwert für Leseverständnis	Retest-Reliabilität aus Testmanual: $r = 0{,}87$ $M = 8{,}61$ ($SD = 3{,}79$)
Problemlösen[2]	komplexes Problemlösen (MicroDyn; Greiff, Wüstenberg, Holt, Goldhammer & Funke, 2013)	7 computerbasierte Items	Wert für Wissensaneignung, geschätzt mittels 2-dim. 2-parametr. IRT-Modell	EAP/PV – Reliabilität für Dimensionen Wissensaneignung 0,79 und Wissensanwendung 0,76
ICT-Nutzung[3]	Nutzungshäufigkeit, adaptiert aus dem PISA Hintergrundfragebogen zur Vertrautheit mit ICT (OECD, 2013)	7 je 4-stufige (0-3) Fragen	Summenwert	Cronbachs alpha: $\alpha = 0{,}73$ $M = 2{,}30$ ($SD = 0{,}50$)

Bemerkung: [1,2,3] In welchen der drei Arbeiten diese Variablen verwendet werden, wird durch die hochgestellten Zahlen markiert.

Aus der Tabelle 3 wird deutlich, dass die Konstrukte zur Erfassung der Personenmerkmale zum Teil eng operationalisiert wurden. Das Verstehen von Text und Grafik wurde mittels eines Lesegeschwindigkeits- und Verständnistests operationalisiert, wodurch beispielsweise die Verarbeitung bildlicher Informationen nicht abgebildet wird. ICT-spezifisches Wissen wurde durch einen Test zur Erfassung des theoretischen Computerwissens ermittelt, der weder Wissen in Bezug auf Normen in einer ICT-Umgebung oder Kriterien der Glaubwürdigkeit enthält. Der Grund für diese Wahl lag in

der Entscheidung für etablierte Verfahren, die computerbasiert umzusetzen und die in die zeitlich begrenzte Erhebung zu integrieren waren.

Tabelle 4. Verwendete Merkmale der ICT-Skills-Items für die empirischen Analysen.

Itemeigenschaft	Indikator	Itemanzahl	Eigenschaften
Leseload[2]	Anzahl der Wörter	64	*M* = 235,9 (*SD* = 250,3; *Min* = 45; *Max* = 1815)
Intrinsische Komplexität (vgl. OECD, 2012, S.50)[2]	Anzahl nicht iterativer Lösungsschritte ohne Navigation zum nächsten Item	64	*M* = 6,0 (*SD* = 3,4; *Min* = 1; *Max*= 16)
Theoretisches erwartetes Lösungsverhalten[1]	Mindestanzahl an Lösungsschritten, inklusive iterativer Schritte und Navigation zum nächsten Item	69	*M* = 9,04 (*SD* = 4,67; *Min* = 3; *Max* = 28)

Bemerkung: [1,2,3] In welchen der drei Arbeiten diese Variablen verwendet werden, wird durch die hochgestellten Zahlen markiert.

Tabelle 5. Verwendete Prozessmerkmale der ICT-Skills-Items für die empirischen Analysen.

Bearbeitungsverhalten	Indikator	Itemanzahl	Eigenschaften[a]
Lösungsschritte[1]	benötigte Anzahl an Lösungsschritten pro Person pro Item	69	*M* = 22 (*SD* = 13; *Min* = 5; *Max* = 74)
Bearbeitungszeit[1]	durchschnittlich benötigte Zeit je Item	69	*M* = 105 Sek. (*SD* = 40; *Min* = 41; *Max* = 241)

Bemerkung: [a] Es wurde zunächst der Mittelwert aller Personen je Item gebildet, dann der Mittelwert über alle 69 Items. [1,2,3] In welchen der drei Arbeiten diese Variablen verwendet werden, wird durch die hochgestellten Zahlen markiert.

4.1.4 Datenanalysen

Die dichotome Bewertung der ICT-Skills-Items erlaubte die Anwendung von Item-Response-Modellen. Hierbei wurde ein eindimensionales ein-parametrisches Item-Response-Modell mit dem R-Paket TAM geschätzt (Kiefer, Robitzsch & Wu, 2016; R Core Team, 2014). Die In- und Outfit-Werte dieses Modells bildeten die Kriterien für die Itemauswahl (Wright & Linacre, 1994; de Ayala, 2013). Bei der Untersuchung der nomothetischen Spanne wurde darüber hinaus ausschließlich das finale Itemset verwendet (vgl. Wenzel et al., 2016), bei dem auch Items mit Differential Item Functioning ausgeschlossen wurden.

Um zu prüfen, inwiefern die Rahmenkonzeption zu Items unterschiedlicher Schwierigkeiten geführt hat (Forschungsfrage 1), wurden die Itemschwierigkeiten mittels des Rasch-Modells ermittelt, um dann Unterschiede in Mittelwert (Varianzanalyse) und Varianz (Homogenität der Varianzen) zu untersuchen. Zur Untersuchung der Konstruktinterpretation (Arbeiten 2 und 3) wurden mit dem R-Paket lme4 (Bates, Maechler, Bolker & Walker, 2014) erklärende Item-Response-Modelle modelliert, bei denen es sich um sogenannte generalisierte lineare Mischmodelle handelt (GLMMs; De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx & Partchev, 2011; Wilson, De Boeck & Carstensen, 2008). Item- und Personenmerkmale (Tabellen 3 und 4) wurden als erklärende Variable für die Wahrscheinlichkeit der Itemlösung aufgenommen und Zufallseffekte für Items, Personen und Schulen modelliert. Während höhere Werte in Item-Response-Modellen die Schwierigkeit anzeigen, bilden höhere Werte in GLMMs die Leichtigkeit ab, da Wahrscheinlichkeiten für die Itemlösung ausgedrückt werden.

4.2 Ergebnisse und Interpretation

Im Folgenden werden die empirischen Ergebnisse der drei Arbeiten separat dargestellt und interpretiert. Detaillierte Darstellungen sind den einzelnen Arbeiten zu entnehmen. Eine Integration der Ergebnisse erfolgt in der Diskussion.

4.2.1 Erprobung der Rahmenkonzeption (Arbeit 1)

Das Ziel der ersten Arbeit bestand zunächst darin, eine theoretische Basis für die Itementwicklung zu schaffen (vgl. 2.2.1) sowie Rationale für die Implementierung verhaltensbasierter Items zu formulieren (vgl. 2.2.2). Obwohl das Hauptaugenmerk auf der Entwicklung der Rahmenkonzeption lag, wurde darüber hinaus untersucht, inwiefern diese beiden konzeptionellen Überlegungen jeweils zu den intendierten Ergebnissen geführt haben.

Hierbei wurde zunächst untersucht (Forschungsfrage 1a), inwiefern die Items der fünf kognitiven ICT-Aufgaben (*Zugriff auf*, *Managen*, *Integrieren*, *Bewerten* und *Erzeugen von Informationen*) einen vergleichbaren Bereich an Itemschwierigkeiten abdecken. Nach dem Levene-Test unterschieden sich die Itemschwierigkeiten der fünf kognitiven ICT-Aufgaben nicht in ihren Varianzen ($F(4, 64) = 0{,}13$, $p = {,}971$). Die Ergebnisse der ANOVA untermauern, dass die kognitiven ICT-Aufgaben vergleichbare mittlere Itemschwierigkeiten hatten ($F(4, 64) = 1{,}12$, $p = {,}356$). Für die Forschungsfrage 1b wurde das theoretisch erwartete Lösungsverhalten (vgl. Tabelle 4) mit dem empirisch gezeigten Lösungsverhalten (vgl. Tabelle 5) verglichen, wobei hierfür das zweite

und dritte Quartil der erfolgten Lösungsschritte herangezogen wurde. In den meisten Items verhielt sich der Großteil der Testteilnehmer/innen so, wie es intendiert war. Es wurden nur wenige Items identifiziert, in denen die angenommenen kognitiven Prozesse vermutlich nicht durchgeführt wurden. In drei der 69 Items vollzogen mehrere Testteilnehmer/innen (>25%) weniger als die zur korrekten Lösung notwendigen Schritte, was in zwei Items durch einen kürzeren falschen Lösungsweg bei gleichzeitig hoher Itemschwierigkeit erklärt werden kann. Im dritten Item benötigt eine falsche Lösung genauso viele Lösungsschritte wie eine richtige, woraus geschlossen werden kann, dass die kognitiven Prozesse bei einem Großteil der Testteilnehmer/innen nicht wie intendiert erfolgten. Dieses Item sollte das Integrieren von Informationen erfordern, wobei ein Großteil der Teilnehmer/innen eine Website auswählte, ohne beide Websites zu betrachten. Diese wurden demnach nicht integriert. Eine Überarbeitung dieses Items könnte an der Spezifizierung der Instruktion ansetzen, damit tatsächlich beide Websites betrachtet werden. Des Weiteren waren zwei Items auffällig, in denen die Testteilnehmer/innen überaus viele Lösungsschritte mit gleichzeitig hoher Variabilität zeigten. Damit einhergehend waren die durchschnittlichen Bearbeitungszeiten für diese beiden Items hoch. Da die Erhebungszeit meist ein knappes Gut ist, sollten diese beiden Items deshalb in ihrer Komplexität reduziert werden.

Anhand dieser beiden ersten Analysen lässt sich zusammenfassend festhalten, dass die Rahmenkonzeption mit den abgeleiteten schwierigkeitsdeterminierenden Merkmalen dazu geeignet war, Items verschiedener Schwierigkeitsausprägung für alle fünf kognitiven ICT-Aufgaben zu entwickeln und diese so zu implementieren, dass sich die Mehrzahl der Testteilnehmer/innen in den meisten Items wie intendiert verhielt.

4.2.2 Evidenzen für Validität basierend auf der nomothetischen Spanne (Arbeit 2)

Das Ziel der zweiten Arbeit bestand darin, Zusammenhänge mit anderen Variablen zu untersuchen, um Evidenzen für die intendierte Konstruktinterpretation zu sammeln. Basierend auf dem in der Arbeit 1 definierten Zusammenspiel der ICT-spezifischen und generischen Fertigkeiten wurden spezifische Erwartungen an das Zusammenspiel dieser drei Fertigkeiten in der Itemlösung formuliert. Neben Zusammenhängen zwischen allen Items wurden aufgrund ihrer Heterogenität die Zusammenhänge auch auf der einzelnen Itemebene betrachtet (vgl. Abschnitt 2.4.2) und die jeweiligen Zusammenhänge durch Itemmerkmale erklärt (Hypothese 2c).

Über alle Items hinweg hatten Problemlösen (β = 0,32; $p < .001$), Leseverstehen (β = 0,17; $p <$,001) und ICT-spezifisches Wissen (β = 0,18; $p <$,001) wie erwartet positive Effekte (Hypothese 2a). Bei der Höhe der Effekte sollte auch die Art der Operationalisierung betrachtet werden, bei der das Problemlösen als einziges Konstrukt durch einen nur am Computer zu administrierenden Test erfasst wird und daher möglicherweise stärker mit den simulationsbasierten ICT-Skills-Items zusammenhängt. Obgleich die Zusammenhänge über Items variierten, worauf die bessere Modellpassung des Modells mit variierenden Zusammenhängen (χ^2 (9) = 19,98; p = ,018) hinweist, zeigte sich das obige auf Testebene gefundene Muster bis auf wenige Items auch auf Itemebene (Hypothese 2b). Hierbei kovariierte das ICT-spezifische Wissen eher negativ mit Itemleichtigkeit (r = -,63) und die generischen Fertigkeiten eher positiv mit Itemleichtigkeit (Problemlösen: r = ,15; Leseverstehen: r = ,50). Für die Lösung einiger weniger sehr leichter Items waren die Zufallseffekte von ICT-spezifischem Wissen äußerst gering. Dies ist problematisch, da die ICT-Spezifität für diese Items nicht gegeben zu sein scheint. Alternativ könnte das für diese Items erforderliche Wissen natürlich auch durch den Wissenstest nicht abgebildet werden. Für einige schwerere Items erhöhte das Leseverstehen nicht die Wahrscheinlichkeit der Lösung, was unproblematisch ist, weil die Problemlösefertigkeiten als weitere generische Fertigkeit dennoch die Wahrscheinlichkeit der Lösung erhöhten. Darüber hinaus könnte die enge Operationalisierung des Konstrukts Verstehen von Text und Grafik nur durch das Leseverstehen dafür verantwortlich sein. Testwerte, die auf diesen Items basieren, können demnach durchaus im Sinne des Konstrukts interpretiert werden, da sie dennoch Fertigkeiten höherer Ordnung erforderten (vgl. Abschnitt 2.2.1). Zusammenhangsanalysen mit Itemmerkmalen (Hypothese 2c) deuteten darauf hin, dass dieser variierende Einfluss für das Problemlösen durch die intrinsische Komplexität der Aufgaben entstand, da ein höherer Einfluss von Problemlösen in komplexeren Items gegeben war (β = 0,11; p = ,016). Dagegen hing ein höherer Einfluss von Leseverstehen entgegen der Erwartung nicht mit einer höheren Textmenge in den Items zusammen (β = 0,02; p = ,670). Somit konnten Itemcharakteristiken für das Problemlösen erklären, warum Problemlöseprozesse beim Lösen von ICT-Skills-Items ablaufen. Hinsichtlich des Leseverstehens wird im Rahmen der Diskussion in der Arbeit 2 eine möglicherweise mangelnde Passung des Indikators (Anzahl der Wörter) für die tatsächlich zu verarbeitende Textmenge diskutiert, weil vermutlich nicht jedes Wort in der ICT-Umgebung auch tatsächlich verarbeitet werden muss, um

die Items zu lösen. Dagegen müssen Lösungsschritte, welche Problemlöseprozesse erfordern sollten, unvermeidlich bei der Itemlösung unternommen werden.

Die Ergebnisse stützten die Interpretation der Testwerte dahingehend, dass sie auf ICT-spezifischem Wissen und den postulierten generischen Fertigkeiten des Leseverstehens und Problemlösens basieren. Die signifikante Variation über Items sowie die Erklärung des Problemlösens durch die intrinsische Komplexität der Items stützt die Angemessenheit des aufgabenzentrierten Ansatzes aufgrund der Heterogenität in den Items. Dass die Textmenge in den Items nicht den Zusammenhang mit dem Leseverstehen erklären konnte und einige Items offenbar kein Wissen erforderten, bedarf jedoch einer weiteren Untersuchung.

4.2.3 Evidenzen für Validität basierend auf der Konstruktrepräsentation (Arbeit 3)

Das Ziel dieser Arbeit bestand darin, den Ansatz der Konstruktrepräsentation auf heterogene Items zu übertragen, um zu untersuchen, ob die Schwierigkeit der Entscheidungen (vgl. Abbildung 2), für die ICT-Skills höherer Ordnung benötigt werden, in den Testwerten repräsentiert ist. Die entwickelten Ansätze, Entscheidungen zu verändern (Change) und zu eliminieren (Eliminate), wurden empirisch angewendet.

Die Veränderungen der in der ersten Arbeit abgeleiteten schwierigkeitsdeterminierenden Itemmerkmale führten zu Schwierigkeitsänderungen in die intendierte Richtung (Hypothese 3a). Wie erwartet waren die Change-Items leichter ($\beta = 0{,}54$; $p < ,001$) beziehungsweise schwerer ($\beta = -0{,}90$; $p < ,001$) als die originalen Items. Zugleich veränderten sich die Zusammenhänge mit ICT-Nutzung und ICT-spezifischem Wissen (Hypothese 3b) nur insofern, als Items, die schwieriger wurden, signifikant stärker mit ICT-spezifischem Wissen zusammenhingen ($\beta = 0{,}21$; $p = ,037$). Dieser stärkere Zusammenhang stellt aber nicht infrage, dass die Items nach wie vor ICT-Skills messen. Die Schwierigkeit einer solchen Entscheidung, von der angenommen wird, dass diese ICT-spezifische Fertigkeiten höherer Ordnung erfordert, wird also in den Testwerten repräsentiert und Testwerte veränderter Items können nach wie vor als ICT-Skills interpretiert werden. Die Eliminierung der Entscheidungen aus den Items führte wie erwartet zu leichteren Items (Hypothese 3c; leichter: $\beta = 1{,}45$; $p < ,001$). Darüber hinaus veränderte sich durch die Eliminierung der Entscheidungen wie erwartet (Hypothese 3d) der Zusammenhang mit der ICT-Nutzung ($\beta = -0{,}24$; $p < ,001$), entgegen der Erwartung aber nicht der mit dem ICT-spezifischen Wissen ($\beta = -0{,}00$; $p = ,961$). Die Entscheidun-

gen waren somit für die Schwierigkeit eines Items, aber noch viel mehr für die Repräsentation des Konstrukts verantwortlich, was die Testwerteinterpretation stützt. Die Frage, weshalb ICT-spezifisches Wissen in den originalen Items und in Eliminate-Items einen vergleichbaren Zusammenhang hat, während gerade die Befunde zu den Change-Items und zur nomothetischen Spanne ebenfalls einen verringerten Zusammenhang erwarten lassen, soll in der Diskussion aufgegriffen werden.

Zusammengefasst können die Testwerte dahingehend interpretiert werden, dass diese ICT-spezifische Fertigkeiten höherer Ordnung erfassen. Experimentelle Untersuchungen konnten zeigen, dass die zu treffenden Entscheidungen für die Schwierigkeit eines Items verantwortlich waren und dass sich sogar das gemessene Konstrukt ändert, wenn nur noch Lösungsschritte und Teilhandlungen durchgeführt werden müssen (vgl. Abbildung 2).

5 DISKUSSION

In der vorliegenden Dissertation wurde im Rahmen von drei eigenständigen Arbeiten untersucht, inwieweit das entwickelte Erhebungskonzept zur Itementwicklung geeignet war (Arbeit 1) und inwiefern die gesammelten Evidenzen die Validität der intendierten Testwerteinterpretation stützen können (Arbeiten 2 und 3). Die Ergebnisse der ersten Arbeit untermauerten zum einen, dass die in der Rahmenkonzeption abgeleiteten schwierigkeitsdeterminierenden Merkmale dazu geeignet waren, Items vergleichbarer Schwierigkeit über alle fünf kognitiven ICT-Aufgaben zu entwickeln. Zum anderen zeigten die Ergebnisse, dass sich die Testteilnehmer/innen in den verhaltensbasiert implementierten Items wie intendiert verhielten. Die erstellte Rahmenkonzeption kann also dahingehend bewertet werden, dass sie zur Entwicklung der Items geeignet ist.

Die Ergebnisse der zweiten und dritten Arbeit zeigten, dass die Testwerte als ICT-spezifische Fertigkeiten höherer Ordnung interpretiert werden können. In der Rahmenkonzeption wurden ICT-Skills als basierend auf dem Zusammenspiel ICT-spezifischer und generischer Fertigkeiten definiert. Dies wurde im Rahmen der zweiten Arbeit zur nomothetischen Spanne empirisch untersucht. Positive Zusammenhänge der Variablen Problemlösen, Leseverstehen und ICT-spezifisches Wissen mit der Lösung von ICT-Skills-Items stützten das postulierte Zusammenspiel. Die Zusammenhänge mit den drei Variablen variierten signifikant über Items, wobei die generischen Fertigkeiten wie erwartet nicht stärker mit schwierigeren Items zusammenhingen. Da eine hohe Schwierigkeit in den Items vorrangig durch ICT-spezifische Aspekte bedingt werden sollte, hätte ein größerer Zusammenhang generischer Fertigkeiten mit schwierigeren Items die Testwerteinterpretation hinsichtlich der ICT-Spezifität infrage gestellt. Da in einigen sehr leichten Items der Zusammenhang mit ICT-spezifischem Wissen äußerst gering war, stellt sich für diese Items die Frage der ICT-Spezifität. Problemlösen hing darüber hinaus stärker mit Items zusammen, die eine höhere intrinsische Komplexität aufwiesen. Leseverstehen war in Items mit einer größeren Textmenge dagegen nicht wichtiger. Dies untermauerte zumindest für das Problemlösen, dass das Erfordernis von Problemlöseprozessen in den Items von den Merkmalen abhing.

In der Arbeit zur Konstruktrepräsentation (Arbeit 3) unterstützten die veränderten Change-Items, dass die Entscheidungen schwierigkeitsdeterminierend waren und die Veränderung der in der ersten Arbeit abgeleiteten Merkmale tatsächlich zu einer verän-

derten Itemschwierigkeit führte, ohne das Konstrukt zu verändern. Zugleich unterstützen die Eliminate-Items, dass die Entscheidungen maßgeblich zur Schwierigkeit der Items beigetragen haben und Items ohne diese Entscheidungsebene ein teilweise anderes Konstrukt erfassten. Es wurde also die Testwerteinterpretation gestützt, dass die Testwerte ICT-spezifische Fertigkeiten höherer Ordnung erfassen und darüber hinaus von generischen und ICT-spezifischen Fertigkeiten abhängen.

Im Folgenden sollen zunächst die betrachteten Evidenzquellen zur Validierung der Testwerteinterpretation zusammengetragen und in einer abschließenden Bewertung diskutiert werden (5.1). Anschließend werden übergreifende Aspekte in den einzelnen Arbeiten integriert (5.2). Zudem wird die Untersuchung kritisch reflektiert (5.3). In einem Ausblick werden Wege der formellen Bildung für ICT-Skills sowie Einsatzmöglichkeiten des ICT-Skills-Tests beschrieben (5.4) sowie ein abschließendes Fazit gezogen (5.5).

5.1 Betrachtete Evidenzquellen

Bei der Entwicklung des Validierungskonzeptes dienten im Wesentlichen die Ansätze von Embretson (1983) als Grundlage. Die nomothetische Spanne wurde in der zweiten Arbeit und die Konstruktrepräsentation in der dritten Arbeit untersucht. Nach den Standards für Validität (AERA, APA & NCME, 2014) werden die Ergebnisse zur nomothetischen Spanne als Evidenz auf der Basis des Zusammenhangs mit anderen Variablen bezeichnet. Die Ergebnisse zur Konstruktrepräsentation werden als Evidenz auf Basis von Antwortprozessen bezeichnet. In der vorliegenden Dissertation wurden Analysen über die Ansätze von Embretson hinaus durchgeführt und auch die Ergebnisse der ersten Arbeit lassen sich als Evidenz für Validität betrachten. Die Ergebnisse der einzelnen Arbeiten sollen deshalb hinsichtlich der betrachteten Evidenzquelle beschrieben werden.

5.1.1 Einordnung der betrachteten Evidenzquellen

Das Ziel der empirischen Beiträge der ersten Arbeit bestand darin, zu untersuchen, inwiefern die Rahmenkonzeption zur Itementwicklung und Implementierung geeignet war. Die erste empirische Analyse in der Arbeit zur Rahmenkonzeption (Forschungsfrage 1a) bezog sich auf die Frage, ob die entwickelten Items einen vergleichbaren Bereich an Itemschwierigkeiten für die fünf kognitiven ICT-Aufgaben abdeckten. Hierbei unterschieden sich die Items der fünf kognitiven Aufgaben weder in Varianz

noch im Mittelwert. Dieses empirische Ergebnis kann als *Evidenz basierend auf der internen Struktur* (AERA, APA & NCME, 2014) verstanden werden, weil die postulierte interne Struktur von vergleichbaren kognitiven ICT-Aufgaben empirisch untermauert wurde. Die Testwerte können also auf der Grundlage dieser fünf kognitiven ICT-Aufgaben interpretiert werden. Im Rahmen der zweiten empirischen Untersuchung (Forschungsfrage 1b) wurde anhand des gezeigten Bearbeitungsverhaltens, nämlich der Anzahl der Lösungsschritte, untersucht, ob die kognitiven Prozesse wie von der Aufgabe intendiert durchgeführt wurden. Der Großteil der Teilnehmer/innen verhielt sich wie intendiert und nur wenige Items müssen überarbeitet werden. Die Ergebnisse der Forschungsfrage können als *Evidenz basierend auf Antwortprozessen* (AERA, APA & NCME, 2014) verstanden werden, weil die Annahmen über die kognitiven Prozesse beim Lösen der Items untermauert wurden. Auf der Grundlage der Anzahl an Interaktionen wurden die für die Itemlösung nötigen Informationen vermutlich tatsächlich eingeholt, wodurch die intendierten Integrations- oder Bewertungsprozesse stattfinden konnten. Die Testwerte können also so interpretiert werden, dass der Großteil der Itemantworten durch die intendierten kognitiven Prozesse zustande gekommen ist.

Zur Untersuchung der nomothetischen Spanne wurden zunächst Zusammenhänge mit anderen Variablen betrachtet. Es zeigten sich positive Zusammenhänge mit ICT-spezifischem Wissen und generischen Fertigkeiten über alle Items (Hypothesen 2a). Dieses Muster der Zusammenhänge zeigte sich bis auf wenige Items auch auf Itemebene (Hypothesen 2b). In den Standards werden solche Analysen zu den *Evidenzen basierend auf Zusammenhängen mit anderen Variablen* (AERA, APA & NCME, 2014) gezählt. Da ICT-spezifisches Wissen sowie Prozesse des Leseverstehens und Problemlöseprozesse als Teil des Konstrukts ICT-Skills verstanden wurden, konnten hier konvergente Evidenzen gesammelt werden. Die Analysen wurden außerdem durch erklärende Variablen (Anzahl an Worten und Lösungsschritten) ergänzt, die die variierenden Zusammenhänge der generischen Variablen erklären sollten (Hypothese 2c). Dieser Analyse lag die Annahme zugrunde, dass diese Aufgabencharakteristiken kognitive Prozesse (Lese- und Problemlöseprozesse) hervorrufen sollten. Diese Fertigkeiten sollten in Aufgaben, welche diese Prozesse in größerem Ausmaß erforderten, wichtiger sein. Itemmerkmale zu quantifizieren stellt einen klassischen Ansatz dar, um Rückschlüsse auf kognitive Prozesse zu ziehen (vgl. Konstruktrepräsentation; Embretson, 1983). Daher liefern die Ergebnisse zusätzlich *Evidenzen basierend auf Antwortprozessen* (AERA, APA & NCME, 2014).

Für die Untersuchung der Konstruktrepräsentation wurden Itemmerkmale manipuliert, die für das Konstrukt zentrale kognitive Prozesse beim Bearbeiten der Items hervorrufen sollten. Ein Verändern (Change-Ansatz; Hypothese 3a) oder Eliminieren (Eliminate-Ansatz; Hypothese 3c) der beim Bearbeiten ablaufenden kognitiven Prozesse änderte die Lösungswahrscheinlichkeit und liefert demnach *Evidenzen basierend auf Antwortprozessen* (AERA, APA & NCME, 2014). Darüber hinaus wurden beide Arten der Manipulation hinsichtlich eines potentiell veränderten Zusammenhangs mit anderen konstrukt-verwandten Variablen untersucht (Hypothesen 3b und 3d). Dadurch sollte geprüft werden, ob sich die Zusammenhänge durch eine Manipulation der zentralen kognitiven Prozesse änderten. Dabei sollten diese durch die Change-Manipulation unverändert bleiben. Durch die Eliminate-Manipulation hingegen sollten sich die Zusammenhänge ändern, wenn die eliminierten kognitiven Prozesse wie angenommen zentral für das zu messende Konstrukt waren. Mit diesen Analysen wurden im Rahmen dieser dritten Arbeit zusätzlich *Evidenzen basierend auf Zusammenhängen mit anderen Variablen* gesammelt (AERA, APA & NCME, 2014). Während bei der Betrachtung der Zusammenhänge mit anderen Variablen klassischerweise neben konvergenten auch diskriminante Evidenzquellen betrachtet werden, also Variablen, mit denen die Testwerte stark oder weniger stark zusammenhängen sollen, wurden in dieser Arbeit nur Variablen betrachtet, die als konvergente Evidenzquellen dienen könnten. Hingegen wurde die Logik der diskriminanten Evidenz auch im Rahmen der Hypothesen 3b und 3d angewandt. Normalerweise werden bei der Untersuchung der diskriminanten Evidenz verschiedene Variablen mit denselben Testwerten in einen Zusammenhang gestellt. In dieser Arbeit blieben die Variablen ICT-Nutzung und ICT-spezifisches Wissen dieselben, aber die Testwerte basierten auf anderen Items. Es wurden zur Sammlung der diskriminanten Evidenzen demnach nicht andere Variablen, sondern andere Items verwendet.

5.1.2 Bewertung der betrachteten Evidenzquellen

Obwohl der Validierungsprozess nie als vollständig abgeschlossen gelten kann, lässt sich auf Basis der Testwerteinterpretationen dennoch eine zusammenfassende Beurteilung der Evidenzen vornehmen (vgl. AERA, APA & NCME, 2014, S. 21f.). Die betrachteten Evidenzquellen fußten in erster Linie auf Zusammenhängen mit anderen Variablen und Antwortprozessen. Aus der intendierten Testwerteinterpretation, Testwerte der ICT-Skills-Items als ICT-spezifische Fertigkeiten höherer Ordnung zu interpretieren, resultierten die zentralen theoretischen Konzeptionen für die Itemerstellung. Diese sind das Handlungsmodell (Abbildung 2), aus dem die Annahmen über zentrale

kognitive Prozesse und Itemmerkmale abgeleitet wurden, sowie das Zusammenspiel ICT-spezifischer und generischer Fertigkeiten. Deshalb sind die betrachteten Evidenzquellen, die eben diese kognitiven Prozesse und den Zusammenhang mit den definierten Konstrukten untersuchen, als angemessen zu bewerten, um die Validität dieser Testwerteinterpretation zu untersuchen.

Darüber hinaus wurden auch verschiedene Validierungsverfahren kombiniert: So wurden bei der Analyse der nomothetischen Spanne, bei der klassischerweise Zusammenhänge mit anderen Variablen betrachtet werden, auch Annahmen über kognitive Prozesse einbezogen und Aufgabenmerkmale quantifiziert. Zusammenhänge mit anderen Variablen wurden in dieser Arbeit dann durch Itemmerkmale erklärt (vgl. Hypothese 2c), wodurch zusätzlich Evidenzen basierend auf Antwortprozessen gesammelt wurden. Bei der Untersuchung der Konstruktrepräsentation, die sich im Wesentlichen auf die Untersuchung der Antwortprozesse bezog, wurden zusätzlich Zusammenhänge mit anderen Variablen betrachtet, indem Zusammenhänge mit solchen Variablen durch Itemmanipulationen verändert wurden (vgl. Hypothese 3d). Neben einer Kombination verschiedener Evidenzquellen wurden in der Arbeit zur Konstruktrepräsentation sowohl experimentelle als auch korrelative Verfahren eingesetzt und kombiniert.

Aufgrund der heterogenen Items wurden zur Betrachtung verschiedener Evidenzquellen traditionelle Verfahren erweitert. In der Arbeit zur nomothetischen Spanne wurden Zusammenhänge mit anderen Variablen auch auf Itemebene betrachtet. In der Arbeit zur Konstruktrepräsentation wurden kognitive Prozesse nicht wie üblich über Itemmerkmale quantifiziert, sondern manipuliert. Hierbei ist die Stärke der Itemmanipulationen zum einen in ihrem experimentellen Charakter, aber auch darin zu sehen, dass die Planung der Validierungsstrategien gemeinsam mit der Itementwicklung stattfand. Itemmerkmale, die zur Entwicklung der Items verwendet wurden, bildeten gleichzeitig den Ansatzpunkt für die Manipulation. Diese Validierungsstrategien wurden für den Einsatz an heterogenen Items entwickelt, können aber auch bei homogenen Items sinnvoll eingesetzt werden. Mit der Betrachtung verschiedener Evidenzquellen und dem Einbezug experimenteller und korrelativer Verfahren sowie deren jeweiliger Kombination konnte die Validität der intendierten Testwerteinterpretation untermauert werden. Auch wenn im Rahmen der Arbeiten einzelne Items identifiziert wurden, die der Überarbeitung bedürfen, so kann über alle Arbeiten hinweg die intendierte Testwerteinterpretation als gestützt betrachtet werden.

5.2 Integration der drei Arbeiten

5.2.1 Konzeptionelle Annahmen

Ein zentraler konzeptioneller Beitrag dieser Dissertation bestand in der Übertragung des Handlungsmodells (vgl. Abbildung 2) auf den ICT-Kontext. Aus diesem Handlungsmodell lassen sich die verwendeten Itemmerkmale aller drei Arbeiten ableiten. Der Wert des Handlungsmodells ist darin zu sehen, dass bei Items, welche relativ komplexe Reaktionen auf verschiedenen Ebenen erfordern – sowohl kognitive Entscheidungen als auch physische Interaktionen – dennoch systematisch beschrieben werden kann, welche Ebene für die Testwerteinterpretation relevant ist. Auf die abgeleiteten Itemmerkmale soll nun eingegangen werden.

Für den Ansatz zur Untersuchung der Konstruktrepräsentation wurden Itemmerkmale manipuliert (Arbeit 3), weil diese in heterogenen Items nicht wie im Ansatz von Embretson (1983) quantifizierbar waren. Zugleich wurden bei der Untersuchung der nomothetischen Spanne als Merkmale über Items die Anzahl der Lösungsschritte sowie die Wortanzahl für alle Items quantifiziert (vgl. Tabelle 3), um zu erklären, inwieweit Problemlöse- und Leseverstehensprozesse in der Itembearbeitung stattfanden. Dies könnte als Widerspruch aufgefasst werden und die Notwendigkeit experimentelle Ansätze zu entwickeln infrage stellen. Der Unterschied beider Arbeiten liegt darin, dass die Anzahl der Lösungsschritte und Worte in der Arbeit zur nomothetischen Spanne ein schwierigkeitsbestimmendes Merkmal nicht für ICT-Skills, sondern für das Problemlösen und das Leseverstehen darstellte. Diese Merkmale wurden also zur Erklärung der generischen und nicht ICT-spezifische Fertigkeiten herangezogen. Nach dem Handlungsmodell (vgl. Abbildung 2) sind aber nicht die Lösungsschritte die wesentlichen Merkmale, welche die Itemschwierigkeit determinieren sollten um Testwerte als ICT-Skills höherer Ordnung interpretieren zu können, sondern die Entscheidungen. Und diese sind in den heterogenen Items weder zwischen Items vergleichbar noch quantifizierbar. Somit bildet die Quantifizierung der Itemmerkmale in der Arbeit zur nomothetischen Spanne keinen Widerspruch zu der Argumentation und experimentellen Manipulation in der Arbeit zur Konstruktrepräsentation.

Die aus dem Handlungsmodell (Abbildung 2) abgeleiteten Lösungsschritte wurden für verschiedene Zwecke unterschiedlich operationalisiert (vgl. Tabellen 4 und 5). Im Rahmen der ersten Arbeit (Forschungsfrage 1b) wurde die theoretische Anzahl der

minimal notwendigen Lösungsschritte zur korrekten Itemlösung (vgl. Tabelle 4) – als Itemmerkmal – mit der Anzahl der empirisch verwendeten Lösungsschritte – einem Prozessmerkmal (vgl. Tabelle 5) – verglichen. In der zweiten Arbeit (Hypothese 2c) wurden die Lösungsschritte zur Erfassung der intrinsischen Komplexität als Itemmerkmal erhoben. Auch wenn die Lösungsschritte zweimal als Itemmerkmal operationalisiert wurden, unterschied sich die Art der Operationalisierung. Während in der ersten Arbeit alle Lösungsschritte gezählt wurden, um Rückschlüsse auf die durchgeführten kognitiven Prozesse ziehen zu können, wurden in der zweiten Arbeit nur die nichtiterativen Lösungsschritte gezählt. Dieses Vorgehen wurde gewählt, weil in der zweiten Arbeit die intrinsische Komplexität erfasst werden sollte, um die Variation von Problemlösen über Items zu erklären. Hierbei wurden nur solche Schritte gezählt, die als neu zu bewerten sind. Dies geschah, weil davon ausgegangen wurde, dass das Öffnen einer E-Mail nicht ein zweites Mal zu problemlösespezifischen Schwierigkeiten führen wird. Da das Problemlösen vor allem für neue, nicht zuvor durchgeführte Lösungsschritte benötigt werden sollte, wurden iterative Schritte nicht gezählt. Demnach können Unterschiede im konzeptionellen Verständnis der Indikatoren diese unterschiedliche Operationalisierung erklären.

Die Vielzahl der Annahmen, die auf Basis der konzeptionellen Arbeiten (vgl. Kapitel 2) abgeleitet werden konnten und somit eine Untersuchung der Testwerteinterpretation zuließen, unterstreicht die Nützlichkeit der konzeptionellen Beiträge dieser Dissertation für ein Konstrukt wie ICT-Skills. Darüber hinaus lassen sich die konzeptionellen Beiträge aber auch auf andere thematische Konstrukte übertragen und können insbesondere für solche Konstrukte, die mit heterogenen Items operationalisiert und simulationsbasiert umgesetzt werden, von Nutzen sein. Prädestiniert für Items mit solchen Herausforderungen sind zum Beispiel Konstrukte, die in Bildungsvergleichsstudien fokussiert werden, zum Beispiel Problemlösen in technologiereichen Umgebungen in PIAAC (OECD, 2012), zur Erfassung von Fertigkeiten des 21. Jahrhunderts (Binkley et al., 2012) oder Schlüsselkompetenzen für lebenslanges Lernen (European Parliament and the Council, 2006). Denn solche Konstrukte basieren meist weniger auf psychologischen Theorien als auf institutionell definierten Wissensdomänen (Watermann & Klieme, 2002).

5.2.2 Empirische Ergebnisse

In allen drei Arbeiten zeigte sich ein konsistentes Ergebnismuster (vgl. Abschnitt 4.2) für die Interpretation der Testwerte im Sinne ICT-spezifischer Fertigkeiten höherer Ordnung. Im Folgenden sollen nun die Ergebnisse der drei Arbeiten verglichen werden. Hierbei wird auf den Zusammenhang der Testwerte mit ICT-spezifischem Wissen (Arbeiten 2 und 3), auf die abgeleiteten Merkmale für die fünf kognitiven ICT-Aufgaben (Arbeiten 1 und 3) sowie auf die Heterogenität des Itempools eingegangen (Arbeiten 1, 2 und 3).

Der Zusammenhang der Items mit ICT-spezifischem Wissen wurde in den Arbeiten zur nomothetischen Spanne (Arbeit 2) sowie in der Arbeit zur Konstruktrepräsentation (Arbeit 3) für jeweils verschiedene Item- (18, 38 und 64 Items) und Personengruppen betrachtet (vgl. Tabelle 2). Die Zusammenhänge waren alle signifikant positiv, wodurch ein stabiler Zusammenhang deutlich wird. Der vergleichsweise kleinere Koeffizient (β = ,18) in der Arbeit zur nomothetischen Spanne (64 Items) im Vergleich zu den originalen Items in der Untersuchung der Konstruktrepräsentation (18 Items: β = ,25; 38 Items: β = ,29) kann durch die zwei zusätzlichen Variablen im Modell – Leseverstehen und Problemlösen – erklärt werden. Gemeinsame Varianzteile mit den anderen Variablen wurden in der Arbeit zur nomothetischen Spanne nicht dem ICT-spezifischen Wissen zugeordnet. Auch die Variation von ICT-spezifischem Wissen über Items in der Arbeit zur nomothetischen Spanne steht im Einklang mit den Befunden in der Arbeit zur Konstruktrepräsentation. In der Arbeit zur nomothetischen Spanne ergab sich ein tendenziell stärkerer Zusammenhang von ICT-spezifischem Wissen für schwierigere Items, obgleich dieser Zusammenhang nicht separat auf Signifikanz geprüft wurde. In der Arbeit zur Konstruktrepräsentation wiesen die in die schwierigere Richtung veränderten Change-Items einen signifikant stärkeren und positiveren Zusammenhang mit ICT-spezifischem Wissen als die originalen Items auf und untermauern somit die Wichtigkeit von ICT-spezifischem Wissen bei schwierigen Items. Die Ergebnisse unterstreichen die ICT-Spezifität der Items, weil schwierigere Items stärker mit ICT-spezifischen Aspekten zusammenhingen, während generische Fertigkeiten eine wichtige, aber nicht hinreichende Voraussetzung zum Lösen von ICT-Skills-Items waren. Entgegen der Erwartung hingen auch Eliminate-Items im Vergleich zu den originalen Items unverändert mit ICT-spezifischem Wissen zusammen, obwohl in diesen die Entscheidungen – als ICT-spezifischer Aspekt der ICT-spezifisches Wissen erfordern sollte – nicht mehr getroffen werden mussten. Wie im Rahmen der dritten Arbeit diskutiert,

erforderten möglicherweise auch Eliminate-Items basales Wissen zum Navigieren, beispielsweise zur Auswahl des korrekten Buttons, weshalb dieser Zusammenhang möglicherweise nach wie vor bestehen blieb. Dieser unveränderte Zusammenhang könnte tiefergehend analysiert werden, um zu verstehen, warum ICT-spezifisches Wissen, das eigentlich in schwierigeren Items wichtig war, in den leichteren Eliminate-Items nach wie vor von Bedeutung war. Davon abgesehen zeigten die Ergebnisse zum Zusammenhang der ICT-Skills-Items mit technischem Wissen trotz unterschiedlicher Anzahlen der analysierten Items und sowohl korrelativer als auch experimenteller Methoden ein konsistentes Ergebnismuster.

Die mithilfe der Rahmenkonzeption in der ersten Arbeit abgeleiteten Itemmerkmale für die fünf kognitiven ICT-Aufgaben wurden für die Itementwicklung verwendet und für den Ansatz der Konstruktrepräsentation in den Items manipuliert (Arbeit 3). Die so entwickelten Items hatten nicht nur vergleichbare Schwierigkeiten in Mittelwert und Varianz für alle fünf kognitiven ICT-Aufgaben (Forschungsfrage 1a), sondern die Itemmerkmale waren auch determinierend für die Schwierigkeit der Items (Hypothese 3a), ohne die Repräsentation des Konstrukts in den Testwerten infrage zu stellen (Hypothese 3b). Die empirischen Ergebnisse stützten neben der Testwerteinterpretation auch die in der Rahmenkonzeption (Arbeit 1) abgeleiteten Itemmerkmale („If the predictions are confirmed empirically, both the theory and the interpretation of scores in terms of the theoretical constructs are supported“; Kane, 2013, S.5).

Dieser Arbeit liegt die Annahme zugrunde, dass bei der Erfassung von ICT-Skills, basierend auf der in dieser Arbeit entwickelten Rahmenkonzeption, heterogene Items entstehen. Daher wurden spezielle Validierungsstrategien erarbeitet (vgl. Abschnitt 2.4). Wie heterogen die resultierenden Items tatsächlich waren, lässt sich anhand der Ergebnisse nachvollziehen. So variierten die Items hinsichtlich der in ihnen enthaltenen Textmenge und der intrinsischen Komplexität (vgl. Tabelle 4). Außerdem benötigten die Items unterschiedlich viel Zeit und die Anzahl tatsächlich gezeigter Lösungsschritte (vgl. Tabelle 5) variierte stark im Vergleich zur theoretisch erforderlichen Anzahl an Lösungsschritten (vgl. Tabelle 4). Neben der Heterogenität in den Anforderungen zeichneten sich die Items demnach auch durch die vielfältigen Möglichkeiten zur Navigation aus, was durchaus ein Ziel der Implementierung war (vgl. Abschnitt 2.2.2). Auch die Ergebnisse zur nomothetischen Spanne stützten diese Heterogenität: Das Modell mit der Variation der ICT-spezifischen und generischen Variablen über Items pass-

te besser als ein Modell, bei dem ein gleicher Zusammenhang für alle Items modelliert wurde (Hypothese 2b). Dieser unterschiedliche Zusammenhang über Items konnte für das Problemlösen sogar durch die variierende Anzahl der erforderlichen nicht-iterativen Lösungsschritte in den Items erklärt werden (Hypothese 2c). Die Annahme der heterogenen Items wurde also empirisch durch die Itemmerkmale, die Prozessmerkmale sowie die Analysen zur nomothetischen Spanne gestützt. Dies untermauert die zu Beginn der Dissertation getroffenen Annahmen hinsichtlich der heterogenen Items (vgl. Abschnitt 2.3).

5.3 Kritische Reflexion

5.3.1 Aspekte der Testentwicklung

Um das Spektrum alltäglicher Aufgaben im ICT-Kontext angemessen abbilden zu können, wurde nach dem Ansatz von Mislevy (2013) zunächst die Domäne analysiert und organisiert. Zudem wurden die hierfür vorliegenden Konzeptionen herangezogen (vgl. Abschnitt 2.2.1). Ausgehend von der intendierten Testwerteinterpretation, die einen kognitiven Fokus hat, und der entwickelten Rahmenkonzeption stellte die Repräsentativität kognitiver ICT-spezifischer Schwierigkeiten für die fünf kognitiven ICT-Aufgaben die oberste Priorität bei der Itementwicklung dar. Es wäre aber auch vorstellbar, die Strukturierung des Inhaltsbereichs nicht anhand dieser kognitiven ICT-Aufgaben sondern anhand verschiedener Software vorzunehmen oder diese auf Basis einer Analyse alltäglich anfallender Aufgaben zu entwickeln. Würde die Arbeit ein anderes Ziel verfolgen, zum Beispiel die Zertifizierung von Fertigkeiten für bestimmte Softwareanwendungen, dann wäre die Repräsentativität der Aufgaben für den zu zertifizierenden Inhaltsbereich wichtiger. Anwendungen wie Textverarbeitung und E-Mail-Postfächer würden dann relevante Facetten für die Itementwicklung darstellen, nicht aber die fünf kognitiven ICT-Aufgaben (siehe zur Übersicht: Ferrari et al., 2012). Wäre die Zielsetzung gegeben, dass die Test-Items in erster Linie alltägliche Aufgaben repräsentieren, unabhängig davon, ob diese nun basale Fertigkeiten oder Fertigkeiten höherer Ordnung erfordern, könnte eine Aufgabenanalyse zu einer ersten Strukturierung des Inhaltsbereiches dienen. Das vorliegende Projekt verfolgte jedoch das Ziel, Fertigkeiten höherer Ordnung zu erfassen. Ausgehend von der intendierten Interpretation der Testwerte kann das aufgabenzentrierte Vorgehen mit Fokus auf die kognitiven ICT-Aufgaben daher als zielführend angesehen werden. Darüber hinaus wurden Jugendliche im Rahmen der Testentwicklung (vgl. Wenzel et al., 2015) in einer Online-

Befragung zu ihren Problemen bei der Benutzung von ICT befragt, um diese Ergebnisse für die Itementwicklung zu verwenden. Ein weiterer Aspekt der inhaltlichen Qualitätssicherung stellte der interne Reviewprozess für die entwickelten Aufgaben dar. Zusammengenommen kann das Vorgehen bei der Itementwicklung demnach als angemessen bewertet werden.

Für die vorliegenden Items wurde ein dichotomes Bewertungsschema gewählt. Dies bedeutet, dass nur bei einer korrekten Lösung aller enthaltenen Entscheidungen die Lösung als korrekt gewertet wurde, bei unvollständiger, falscher und auch teilweise falscher Lösung hingegen als inkorrekt. Dabei ist zu beachten, dass die Aufgaben sehr komplex sind, mehrere Entscheidungen, Teilhandlungen und Lösungsschritte enthalten und sich aufgrund der Heterogenität in diesen Merkmalen sehr unterscheiden. Demnach könnte man sich fragen, inwiefern ein solches Bewertungsschema angemessen ist oder ob auch Punkte für eine teilweise richtige Lösung vergeben werden sollten. In dem Beispiel-Item in der Arbeit zur Konstruktrepräsentation gilt es, fünf Entscheidungen zu treffen, das heißt für jede der fünf E-Mails zu entscheiden, ob diese weitergeleitet werden muss (vgl. Abbildung 1, Arbeit 3). Nur wenn alle Entscheidungen korrekt getroffen wurden wird das Item als richtig gelöst bewertet. Sobald eine E-Mail fälschlicherweise weitergeleitet oder versäumt wurde, die relevante E-Mail weiterzuleiten, wird die Lösung als inkorrekt gewertet. Der kritische und ICT-spezifische Aspekt, der auch den Ansatzpunkt für die Manipulation zum Zweck der Validierung in diesem Item bildete, ist die Bewertung der dritten E-Mail. Diese E-Mail enthält eine Falschmeldung und sollte nicht weitergeleitet werden. Würde man hier Punkte für eine teilweise korrekte Lösung vergeben, so wäre es selbst dann zentral, dass diese E-Mail nicht weitergeleitet wurde. Im Sinne von ICT-Skills wäre es weniger relevant, welche Entscheidungen für die anderen E-Mails getroffen wurden, da hierfür in erster Linie Leseverstehen benötigt wird. Problematisch in einem solchen Item bei einer Bewertung mit Teilpunkten ist, dass eine teilweise richtige Lösung das nicht-Weiterleiten der dritten E-Mail erfordert, also keine Reaktion. Es bräuchte also ein faires Kriterium um zielführendes Verhalten in der Aufgabe zu erfassen, damit davon ausgegangen werden kann, dass die Entscheidung für die dritte E-Mail tatsächlich getroffen wurde. Eine Kombination aus diesem zielführenden Verhalten und dem nicht-Weiterleiten der dritten E-Mail könnte mit Teilpunkten belohnt werden. Ein solches Kriterium für zielführendes Verhalten könnte das Weiterleiten irgendeiner E-Mail sein, oder aber das Lesen aller E-Mails. Beide Kriterien wären nicht zufriedenstellend: Nach dem ersten Kriterium würde eine Person, die eine

falsche E-Mail weitergeleitet hat und sich möglicherweise auch nur diese eine angeschaut hat, tatsächlich besser abschneiden als eine Person, die zwar alle E-Mails angeschaut, aber keine weitergeleitet hat. Nach dem zweiten Kriterium würde man das Verhalten von Personen, die zum Beispiel E-Mails bereits aufgrund von auffälligen Adressaten oder Betreffzeilen bewusst nicht lesen, fälschlicherweise als nicht ICT-kompetent bewerten. Ein Bewertungsschema mit Teilpunkten könnte für bestimmte Items – wie für das aufgeführte Beispiel – demnach nicht unbedingt fairer sein. Aus ICT-Perspektive könnte hingegen argumentiert werden, dass bereits eine teilweise richtige Lösung im ICT-Kontext natürlich schon negative Konsequenzen haben kann, wenn zum Beispiel soziale Konventionen nicht beachtet werden bei der Frage, welche E-Mail weiterzuleiten ist und welche nicht. Insofern ist das dichotome Bewertungsschema nicht ungeeignet, obwohl auch hier nicht ausgeschlossen werden kann, dass eine Person die Aufgabe vielleicht nicht wie intendiert bearbeitet hat. Liest eine Person zum Beispiel zufällig nur die vierte E-Mail und leitet diese weiter, während sie die E-Mail mit der Falschmeldung ungelesen lässt, würde das Item als richtig gelöst bewertet werden. Die intendierten kognitiven Prozesse, die E-Mail mit der Falschmeldung richtig zu beurteilen, würden in diesem Fall nicht vollzogen werden. Ob die intendierten kognitiven Prozesse tatsächlich durchgeführt wurden könnte jedoch separat überprüft werden, wie es in Forschungsfrage 1b gemacht wurde, und muss nicht zwangsweise durch das Bewertungsschema implementiert werden. Anstatt die Items detaillierter zu bewerten, sollte eher sichergestellt werden, dass sich die Testteilnehmer/innen durch eine sinnvolle und eindeutige Instruktion so verhalten, wie es intendiert ist.

Punkte für eine teilweise richtige Lösung wären möglicherweise dann sinnvoller, wenn es mehrere Wege gibt, das Item zu lösen, und wenn dabei ein bestimmtes Vorgehen als kompetenter einzustufen ist. In den implementierten Aufgaben ist es tatsächlich so, dass es nur einen Lösungsweg gibt. Es sind aber auch Items vorstellbar, bei denen es nicht allein um die Lösung geht, sondern vielmehr um den Lösungsweg (vgl. Engelhardt, Goldhammer, Naumann & Hartig, 2016, März). Der Computer erlaubt es, Aufgaben auf unterschiedlichen Wegen zu lösen und manche Wege sind zeitlich gesehen effektiver oder auch weniger fehleranfällig (Bhavnani, Peck & Reif, 2008). Ein „smarterer“ Lösungsweg würde iterative Schritte an den Computer delegieren, anstatt diese selbst iterativ durchzuführen. Für solche Items, die unterschiedlichen Lösungswege erlauben, wäre eine abgestufte Punktevergabe auf jeden Fall zu unterstützen.

Der entwickelte ICT-Skills-Test ist ein simulationsbasierter standardisierter Leistungstest, der in größeren Erhebungsstudien zum Einsatz kommen kann. Da Konstruktionen von Simulationsumgebungen sehr aufwändig sind und es gleichzeitig auch denkbar ist, Testpersonen Aufgaben in reeller Software bearbeiten zu lassen (vgl. „authentic tasks“, Parshall, Spray, Kalohn & Davey, 2002; Siddiq et al., 2016), werden im Folgenden Eigenschaften beider Verfahren beschrieben, um die Wahl für die Simulationsumgebung zu reflektieren. Ein *erster* Unterschied zwischen den beiden Verfahren besteht darin, dass sich Testpersonen in Simulationsumgebungen nur innerhalb des gesteckten Rahmens bewegen können, während in reeller Software keine Grenzen an verfügbaren Funktionen in Programmen oder an Navigationsmöglichkeiten existieren. Die Bewegungseinschränkungen in Simulationsumgebungen können als nachteilig empfunden werden, weil einer Person nicht alle Lösungswege wie im Alltag zur Verfügung stehen. In einer reellen Aufgabe könnten Personen fehlendes Wissen (z.B. über verschiedene Speicherformate) auch durch eine Internetrecherche ausgleichen. Die Einschränkungen in Simulationsumgebung können demnach auch als Vorteil verstanden werden, weil eine falsche Lösung beispielsweise auf fehlendes Wissen über Speicherformate zurückgeführt werden kann, wenn Personen keine Internetrecherche durchführen können. Ob solche kompensatorischen Möglichkeiten in den Testwerten enthalten sein sollen oder nicht, kann einen Grund für die Wahl der Umgebung bilden. Darüber hinaus gibt es natürlich auch in reeller Software die Möglichkeit, Bearbeitungswege zu restringieren, zum Beispiel wenn es keinen Internetzugang gibt oder bestimmte Software nicht verfügbar ist. Ein *zweiter* Unterschied besteht darin, dass bei der Verwendung reeller Software Personen, die zu Hause andere E-Mail-Postfächer oder Textverarbeitungssoftware nutzen, an einem Testcomputer benachteiligt werden. Würden alle Nutzer die Aufgabe mit ihnen vertrauter – also unterschiedlicher – Software bearbeiten, stellt sich die Frage der Vergleichbarkeit der Testwerte. Aufgaben in einer Simulationsumgebung können hingegen so implementiert werden, dass sie keine bestimmte Software kopieren, sondern von existierender Software abstrahiert sind. Dagegen erfordern Simulationsumgebungen Transferprozesse, zum Beispiel weil zunächst der richtige Button zum Versenden einer E-Mail gesucht werden muss. Solche Transferprozesse, Wissen aus anderen Situationen auf eine neue Situation zu übertragen, könnten durchaus als Teil des Konstrukts ICT-Skills angesehen werden. Ob Transferprozesse in den Testwerten repräsentiert sein sollen kann die Wahl für oder gegen eine Simulationsumgebung beeinflussen. Darüber hinaus könnten existierende Bibliothekssysteme oder interne

Firmensysteme hier einen Mittelweg darstellen, da diese reelle Umgebungen darstellen, bei denen keine Vorteile für bestimmte Nutzergruppen zu erwarten sind. Inwiefern Testwerte aus solchen Systemen im Sinne alltäglicher Fertigkeiten interpretiert werden können, müsste eigens geprüft werden, könnte aber für bestimmte Erhebungsziele, etwa für berufliche Einstellungstests, eine akzeptable Lösung darstellen. *Drittens* ist es denkbar, dass sich Bearbeitungswege von Aufgaben in reeller Umgebung im Laufe der Zeit verändern, zum Beispiel, weil Internetseiten einem fortwährenden Wandel unterliegen oder weil es neue Softwareversionen gibt. Die Vergleichbarkeit von Testwerten für Aufgaben in reeller Umgebung, die keinen Internetzugriff erfordern, könnte durch eine regelmäßige Überprüfung gewährleistet werden. Dahingegen wäre eine Vergleichbarkeit von Testwerten für Aufgaben, die zeitlich verzögert gestellt werden und die Nutzung von Internet erfordern, nicht gegeben. Auch Simulationsumgebungen müssen von Zeit zu Zeit modernisiert werden, stellen dagegen aber eine kontrollierbare Umgebung dar. *Viertens* stellt sich die Frage der ökonomischen Bewertung. Das Verhalten in einer Simulationsumgebung ist mithilfe vordefinierter Regeln automatisch bewertbar. Es kann also unmittelbare Rückmeldung erfolgen. Zudem lassen sich die Items dadurch auch adaptiv administrieren. In reeller Software könnte eine solche automatische und unmittelbare Bewertung erfolgen, wenn die Antwortabgabe auf einem Antwortbogen außerhalb der Aufgabe vorgenommen wird. In Simulationsumgebungen ist dank der limitierten Bearbeitungswege neben der Bewertung der Aufgabe auch Bearbeitungsverhalten sehr ökonomisch zu erfassen, zum Beispiel ob eine bestimmte Information eingeholt wurde. In reeller Umgebung dagegen kann eine Kodierung von Verhalten durch die Anzahl möglicher Bearbeitungswege um einiges komplexer sein. *Zusammengenommen* gilt es, die Vor- und Nachteile im Hinblick auf die Zielsetzung des Testverfahrens abzuwägen. Während die Bewertung des gezeigten Verhaltens leichter und ökonomischer in einer Simulationsumgebung umzusetzen ist, erfordert eine von reeller Software abstrahierte Simulationsumgebung gleichzeitig Transferprozesse und erlaubt nur bestimmte – implementierte – Bearbeitungswege. Fehlendes Wissen kann dann nicht durch andere Strategien ausgeglichen werden, die auch als Teil des Konstrukts angesehen werden könnten. Für die Ziele des vorliegenden Instruments, das potentiell auch in Large-Scale-Assessments einsetzbar und adaptiv administrierbar sein soll[2], ist die Wahl der Simulationsumgebung jedoch als sinnvoll anzusehen.

[2] Die Entwicklung des ICT-Skills-Tests erfolgte im Projekt CavE-ICT-PISA, das die Entwick-

5.3.2 Datenanalysen

In der zweiten und dritten Arbeit wurden GLMMs, generalisierte lineare Mischmodelle, verwendet. Diese erlauben es, neben festen Effekten auch Zusammenhänge auf Itemebene – Zufallseffekte – zu modellieren. Zusammenhänge dürfen dann über Items variieren, was die Datenstruktur gerade bei heterogenen Items besser abbilden mag. Es kann darüber hinaus überprüft werden, ob diese Zufallseffekte zu einer signifikant besseren Modellpassung führen, während die Zusammenhänge für einzelne Items nicht auf Signifikanz geprüft werden. In der Arbeit zur nomothetischen Spanne wurde das Muster dieser Zufallseffekte betrachtet, hinsichtlich der Zusammenhänge mit Itemschwierigkeit sowie der Betrachtung von Zusammenhängen die für einzelne Items nahe null waren. An dieser Stelle soll darauf hingewiesen werden, dass auf Basis dieser Zufallseffekte keine Items ausgeschlossen wurden. Diese Zufallseffekte stellten eher eine grafische Inspektion dar als ein striktes Kriterium zur Itemauswahl. Ein solches grafisch inspizierendes Vorgehen wurde auch im Rahmen der dritten Arbeit empfohlen, um einen Eindruck davon zu erhalten, in welchen Items die Veränderung der Merkmale einen großen oder kleinen Effekt auf die Schwierigkeit hatte. Diese Möglichkeit der Inspektion ist gerade in heterogenen Itemsets hilfreich, da hier eine stärkere Variation von Effekten zu erwarten ist. Aus diesem Grund können GLMMs eine sinnvolle Analysemethode für heterogene Items sein.

Alle Modelle der methodischen Analysen basieren auf eindimensionalen und ein-parametrischen Modellen der Item-Response-Theorie. Basierend auf den fünf kognitiven Aufgaben (*Zugreifen auf*, *Managen*, *Integrieren*, *Bewerten* und *Erzeugen von Informationen*) wäre auch ein fünfdimensionales Modell plausibel gewesen. Im Rahmen der Skalierungsarbeiten des Tests, die nicht Teil dieser Arbeit waren, wurde letztlich ein eindimensionales Modell präferiert (vgl. Wenzel et al., 2015, S.10), welches auch in dieser Arbeit verwendet wurde. Auch die Analysen der ICILS-Studie fußen – obgleich auch ein zweidimensionales, auf der Rahmenkonzeption basierendes Modell plausibel gewesen wäre – letztlich auf eindimensionalen Modellen (Fraillon, Ainley, Schulz, Friedman & Gebhardt, 2014). Für die entwickelten Items kann deshalb die Frage aufkommen, wie eindeutig diese Items tatsächlich auf die kognitiven ICT-Aufgaben gepasst haben. Obwohl die Passung der Items mittels verschiedener Review-Prozesse dis-

lung eines computergestützten, adaptiven und verhaltensnahen Instruments zur Erfassung Informations- und Kommunikationstechnologiebezogener Fähigkeiten (ICT-Skills) zum Ziel hatte, das z.B. in Large-Scale-Assessment-Studien wie PISA eingesetzt werden kann.

kutiert wurde (vgl. Wenzel et al., 2015) ist es denkbar, dass die Items, wie das vermutlich auch bei alltagstypischen Aufgaben der Fall ist, ebenso Aspekte anderer kognitiven Aufgaben enthalten. So könnte das Integrieren von Informationen außerdem Bewertungsaspekte enthalten, zum Beispiel hinsichtlich der Relevanz einer zu integrierenden Informationseinheit. Die bessere Passung des eindimensionalen Modells kann demnach auch durch nicht sortenreine Items bedingt sein. Da alle Items – unabhängig von der Zugehörigkeit zu den fünf kognitiven ICT-Aufgaben – auf Problemlösen, Leseverstehen und ICT-spezifischem Wissen basieren sollten, also vergleichbare Fertigkeiten für Items von allen ICT-Aufgaben erforderlich sind, widerspricht ein eindimensionales Modell nicht der intendierten Testwerteinterpretation und verwendet eine andere Studien (vgl. ICILS) ähnelnde Modellierung von ICT-Skills.

5.4 Ausblick

5.4.1 Wege der formellen Bildung

Der Fokus dieser Arbeit lag in erster Linie auf der Erstellung eines Erhebungs- und Validierungskonzepts, um ICT-Skills im Rahmen groß angelegter Studien für 15-jährige Schülerinnen und Schüler erfassbar zu machen. Demnach lag der Fokus auf Personen, die noch zur Schule gehen, weshalb die Frage naheliegt, welche Implikationen aus den Ergebnissen dieser Arbeit für die Praxis abgeleitet werden können. Obwohl der Umgang mit Informations- und Kommunikationstechnologien kein klassisches Schulfach darstellt, werden ICT-Skills dennoch zu den Schlüsselkompetenzen des 21. Jahrhunderts und für ein lebenslanges Lernen gezählt (European Parliament and the Council, 2006; Binkley et al., 2012). Medienbildung ist zum Beispiel auch im Lehrplan von Baden-Württemberg im Jahr 2016 eine von sechs Leitperspektiven[3] und soll fächerintegriert eingebunden werden. Der kompetente Umgang mit Medien und Technologien stellt also durchaus ein relevantes Bildungsziel dar.

Die Fragestellungen der vorliegenden Dissertation betrafen im Wesentlichen die Validität der Testwerteinterpretation. Die Ergebnisse zur nomothetischen Spanne stützten die Annahme, dass Personen mit höheren Problemlösefertigkeiten, besserem Leseverstehen und mehr Wissen eine höhere Wahrscheinlichkeit hatten, Items zu lösen. Diese Variablen klärten einen beträchtlichen Anteil der Varianz zwischen Personen auf (35,4%). Sollen Schülerinnen und Schüler im Rahmen der Schule auf solche Herausfor-

[3] http://www.bildungsplaene-bw.de/,Lde/LS/BP2016BW/ALLG/LP letzter Aufruf 31.8.2017

derungen vorbereitet werden, wie sie im Fokus dieser Arbeit standen, so könnten die hier untersuchten generischen und ICT-spezifischen Fertigkeiten gefördert werden. Leseverstehen wird bereits in der Schule thematisiert, nicht nur im Fach Deutsch selbst, sondern auch durch das Lesen von Texten in anderen Fächern und sogar in Mathematik in Textaufgaben. Auch wenn Problemlösen kein Unterrichtsfach ist, so wird es dennoch in anderen Fächern gefördert, zum Beispiel beim Experimentieren und Durchführen von Versuchen in naturwissenschaftlichen Fächern. Hinsichtlich des Wissens kann nun gefragt werden, welches Wissen relevant ist und wie dieses vermittelt werden sollte. Da sich die vorliegende Dissertation nicht mit Wissensaneignung in ICT-Umgebung beschäftigt hat, sollen im Folgenden nur einige Denkanstöße gegeben werden.

Ein wichtiger Aspekt an ICT-Umgebungen ist, dass sich diese ständig weiterentwickeln und einst erlernte spezifische Funktionen in spezieller Software zum Studien- oder Ausbildungsbeginn eventuell bereits veraltet oder an anderer Stelle zu finden sind. Genau genommen können Nutzer/innen über den Umgang mit Informations-und Kommunikationstechnologien nie so viel gelernt haben, um dauerhaft als kompetent gelten zu können. Es ist also wichtig, Menschen, die in Zukunft kompetent mit ICT umgehen sollen, neben konkretem Wissen über ICT auch Konzepte zu vermitteln, die dabei helfen, Gelerntes selbständig auf neue Situationen zu übertragen und Probleme selbständig zu beheben. Dies könnte erreicht werden, wenn Wissen nicht in einer Weise erworben wird, in der es an eine bestimmte Version gebunden ist. Dadurch ließe es sich in abstrahierter Form auf neue Situationen übertragen (vgl. Carroll & Rosson, 1987; Day & Goldstone, 2012; Singley & Anderson, 1985). So kann es wichtiger sein, Sinn und Funktionsweisen von Verzeichnissen (z.B. eines Inhaltsverzeichnisses) zu verstehen, als zu wissen, wo genau eine bestimmte Funktion in der aktuellen Textverarbeitungssoftware zu finden ist. Bhavnani et al. (2008) beschreiben zum Beispiel verschiedene Funktionsweisen, die einen Computer zu einem mächtigen Werkzeug machen. Diese Funktionsweisen zu nutzen bedeutet, effektive Strategien einzusetzen, die zeitsparend und fehlervermeidend sind. Ein Beispiel hierfür ist die Vermeidung iterativer Schritte, indem diese an den Computer delegiert werden („Iteration“). Das gleichzeitige Ändern aller Fußzeilen auf Präsentationsfolien mithilfe des Menüs würde iterative Schritte vermeiden, weil die Fußzeilen nicht manuell auf jeder Folie einzeln angepasst werden müssen.

Solche Strategien müssen sich ICT-Nutzer aufgrund der ständigen technologischen Weiterentwicklung meist selbständig erarbeiten oder an neue Software anpassen, was die Voraussetzung dafür ist, um auch in Zukunft kompetent mit ICT umzugehen. Ein Verständnis für derartige Funktionsweisen zu haben, kann beim Aneignen und Anwenden solcher „smarten" Strategien in konkreter Software helfen und Ziel der Vermittlung von ICT-Skills sein. Sollten Fertigkeiten abstrahiert von bestimmten Programmen geschult werden, ist es auch denkbar, Unterrichtseinheiten anhand der kognitiven ICT-Aufgaben zu strukturieren, wie sie auch im Rahmen dieser Arbeit zur Erfassung kognitiver ICT-Fertigkeiten verwendet wurden. Themen könnten dann die Beschaffung von Informationen (Zugriff auf Informationen) oder Glaubwürdigkeitskriterien (Bewertung von Informationen) sein.

Da ICT-Skills für das Meistern alltäglicher Aufgaben relevant sind, ist es sinnvoll, solche Fertigkeiten auch im Bildungsbereich zu adressieren. Da generische Fertigkeiten durchaus im aktuellen Curriculum Berücksichtigung finden, stellt sich in erster Linie die Frage, wie das nötige ICT-spezifische Wissen vermittelt werden sollte, sodass Nutzer/innen trotz technologischer Weiterentwicklung kompetent sind. Dieses Wissen abstrahiert von konkreter Software zu vermitteln, ist ein möglicher Weg, der hier aufgezeigt wurde. Weitere Forschungsbemühungen im Bereich von ICT-Skills könnten daran ansetzen, den Prozess der Aneignung sowie die Rolle von Konzepten bei der Aneignung neuen Wissens im ICT-Kontext zu untersuchen.

5.4.2 Einsatz- und Weiterentwicklungsmöglichkeiten des ICT-Skills-Tests

Der entwickelte ICT-Skills-Test spiegelt alltagstypische Aufgaben wider. Er bezieht sich weder auf ein klassisches Schulfach noch bildet er ein Level ab, über das Schüler/innen nach einem bestimmten Schuljahr verfügen sollten. Demnach ist es denkbar, die Items auch für andere Personengruppen einzusetzen, zum Beispiel für Studenten oder Berufstätige. Hierbei muss zunächst beachtet werden, dass sich die im Rahmen dieser Arbeit gesammelten Evidenzen für Validität nur auf die intendierte Testwerteinterpretation beziehen und dass Evidenzen für diese Interpretation anhand einer spezifischen Stichprobe – 15-jährige Schülerinnen und Schüler – gesammelt wurden. In Studien, die ICT-spezifische Fertigkeiten in verschiedenen Altersgruppen untersucht haben, wurde gezeigt, dass jüngere Nutzer/innen zwar bessere photovisuelle Fertigkeiten sowie eine bessere Orientierung haben, aber dafür in der Bewertung von Informationen schlechter sind (Eshet-Alkalai & Amichai-Hamburger, 2004; Eshet-Alkalai & Chajut,

2010; Lorenzen, 2001; van Deursen & van Dijk, 2009). Demnach könnten bestimmte Items für andere Personengruppen leichter oder schwerer sein, zum Beispiel aufgrund von erforderlichem Wissen. Weitere Interpretationen sowie eine Anwendung auf andere Personengruppen erfordern die Sammlung neuer Evidenzen. Da sich die Validitätsbemühungen dieser Dissertation auf die Validierung der Konstruktinterpretation bezogen, wurden zum Beispiel keine kriterienorientierten Interpretationen untersucht. Sollten die ICT-Skills-Items als Auswahlinstrument dienen, müsste die Validität einer kriterienorientierten Interpretation untersucht werden, zum Beispiel, ob Personen mit höheren Testwerten einen größeren Erfolg bei der Aneignung neuer Computersysteme haben.

Soll der entwickelte ICT-Skills-Test eingesetzt werden, so müsste der gegenwärtig auf 64 Items basierende Itempool gekürzt werden, da bei einer durchschnittlichen Bearbeitungszeit von 1-2 Minuten die Testzeit zu lang wäre. Basierend auf 25 ICT-Skills-Items wurde deshalb eine Kurztestskala sowie ein Algorithmus zur adaptiven Vorgabe entwickelt (vgl. Wenzel et al., 2016). Darüber hinaus ist es auch denkbar, die entwickelten Change-Items zu verwenden, wenn Items mit einer höheren oder geringeren Itemschwierigkeit erforderlich sind. Denn diese sind Varianten der originalen Items, die leichter oder schwerer sind, ohne dass die Testwerte ein anderes Konstrukt repräsentieren. Für Interpretationen, die auf neuen Testzusammenstellungen basieren, müssten natürlich erneut Evidenzen gesammelt werden.

Ebenso ist es naheliegend, die ICT-Skills-Items als Kontrollvariable für computerbasierte Erhebungen einzusetzen, um ermitteln zu können, ob sich durch höhere ICT-Skills Vorteile ergeben. Wenn das zu interessierende Konstrukt (zum Beispiel Bewertungskompetenz oder Problemlösen) jedoch auf kognitiven Fertigkeiten basiert und beispielsweise Lesen und Problemlösen als Teil dieses Konstrukts angesehen werden, dann bestünde die Gefahr, konstrukt-relevante Varianz fälschlicherweise aus den Analysen auszuschließen. Die entstandenen Eliminate-Items könnten in solchen Fällen möglicherweise eher als Kontrollvariable geeignet sein, da in diesen Items die Entscheidungen – die Fertigkeiten höherer Ordnung erfordern sollten – eliminiert wurden und die erforderlichen Fertigkeiten zum Lösen der Items basaler wären als bei den ICT-Skills-Items. Ebenso könnte ein Test, der basale Fertigkeiten im ICT-Kontext abbildet, besser geeignet sein (vgl. Basic Computer Skills; Goldhammer et al., 2013).

Davon abgesehen kann eine Simulationsumgebung, wie sie in dieser Arbeit verwendet wurde, auch dazu nützlich sein, Wissen in abstrahierter Form zu generieren,

um das Lernen stärker auf das Verstehen von Konzepten zu lenken, anstatt konkretes Wissen in aktueller Software zu erzeugen. Demnach könnten solche simulationsbasierten Umgebungen nicht nur zu Test-, sondern auch zu Trainingszwecken sinnvoll sein. Die Items könnten so weiterentwickelt werden, dass Lernende je nach Art des Fehlers mit zusätzlichen Informationen ausgestattet werden, zum Beispiel zu Falschmeldungen (vgl. Formatives Assessment, Black & William, 1998). Das Gelernte könnte in solchen Umgebungen dann gleich verhaltensbasiert in der Aufgabe angewendet werden. Ein formativer Aufgabencharakter könnte auch dazu genutzt werden, Schüler/-innen zu „smarteren" Bearbeitungswegen zu motivieren, also die Möglichkeiten von Computern zu nutzen und weitgehend effektiv zu arbeiten.

Eng mit dem Thema der Weiterverwertung der Items ist die Frage verbunden, wie schnell die Konzeption und die Items überholt sind. Durch die ständige Weiterentwicklung von Technologien ist es unvermeidlich, dass solche simulationsbasierten Items nicht ohne Überarbeitung über Jahre oder Jahrzehnte hinweg eingesetzt werden können. Neben der geschaffenen Simulationsumgebung, die bereits in wenigen Jahren veraltet aussehen wird, werden sich auch Inhalte ändern. Somit ist die Lebensdauer eines solchen Tests weitaus geringer als zum Beispiel von Tests zur Intelligenzdiagnostik (vgl. Raven, 2000). Für die vorliegende Arbeit muss hier zunächst zwischen der Rahmenkonzeption und den entwickelten Items unterschieden werden. Der kognitive Fokus der Rahmenkonzeption auf den Inhaltsbereich ICT rückt kognitive Aufgaben wie den *Zugriff auf* oder das *Erzeugen von Informationen* in den Mittelpunkt. Die Schwierigkeit beim Zugriff auf Informationen kann beispielsweise darin bestehen, passende Suchbegriffe auszuwählen und diese iterativ anzupassen. Dabei ist es weniger wichtig, wie die Suchmaske aussieht und ob dieser Begriff nun eingetippt oder über eine Sprachfunktion eingegeben wird, sondern vielmehr, ob die Kombination der gewählten Suchbegriffe zielführend ist. Die Konzentration auf kognitive Schwierigkeiten in den Items ist demnach ein Aspekt, der ein allzu schnelles Veralten abfedern kann. Nicht zu leugnen ist jedoch die abnehmende Authentizität der Aufgaben, wenn Oberflächen veraltet aussehen, das Veralten von Funktionen, die in den Items implementiert sind (z.B. klassische Menüs oder Ribbons), oder die Änderung von Wissensbeständen. So haben sich Speichermedien in den letzten Jahrzehnten stark verändert und die lange Zeit bewährte Diskette wurde von CD-ROMs und USB-Sticks ersetzt. Solche Veränderungen gelten für simulationsbasierte Items genauso wie zum Beispiel reine Wissenstests im Multiple-Choice-Format. Allerdings sind die durch Veraltung entstehenden Kosten wegen des

hohen Entwicklungsaufwands bei simulationsbasierten Items höher. Aufgrund der Simulationsumgebung, die von konkreten Applikationen abstrahiert, veralten Items aber nicht so schnell, wie es beispielsweise der Fall wäre, wenn Software aus dem Alltag eins zu eins simuliert würde. So mussten in den Aufgaben eher generelle Konzepte – beispielsweise über Speicherformate – angewendet werden, als dass tatsächlich existierende Speicherformate ausgewählt werden mussten. Ebenso waren Menüs nicht an derselben Stelle wie in speziellen Softwareversionen zu finden. Solche Aspekte beugen dem allzu schnellen Veralten vor, können dieses aber natürlich nicht verhindern. Auch wenn die kognitiven Schwierigkeiten in den Items nach wie vor relevant sein mögen, müssen die Items in kürzeren Zyklen überarbeitet und angepasst werden, als dies beispielsweise bei anderen Tests der Fall ist (vgl. Raven, 2000).

5.5 Schlussfolgerung

In dieser Arbeit wurde ein Erhebungs- und Validierungskonzept zur Erfassung kognitiver Fertigkeiten im Umgang mit ICT beschrieben. Mit ICT-Skills wurde ein Inhaltsbereich in den Blick genommen, der sich von klassischen Konstrukten der psychologischen Leistungsdiagnostik unterscheidet (z.B. Intelligenzdiagnostik), weshalb für die Entwicklung des Erhebungs- und Validierungskonzepts konzeptionelle Arbeit notwendig war. Die Herausforderung dieser Arbeit, etwa die fehlende theoretische Grundlage für die Itementwickung oder der sehr heterogene Itempool für die Validierungsstrategien, treten nicht allein bei der Erfassung von ICT-Skills auf, sondern sind durchaus auch für die Erfassung weiterer Fertigkeiten des 21. Jahrhunderts (Binkley et al., 2012) und für andere simulationsbasierte Erhebungen relevant. Mit dieser Arbeit konnte gezeigt werden, dass sich auch Konstrukte, die auf institutionell definierten Wissensdomänen fußen (Watermann & Klieme, 2002), an den Vorgehensweisen der psychologischen Leistungsdiagnostik orientieren und einen Fokus auf die kognitiven Prozesse richten können. Darüber hinaus wurde untermauert, dass die Rahmenkonzeption für die Itementwicklung geeignet war und dass die Testwerte des entwickelten ICT-Skills-Tests im Sinne ICT-spezifischer Fertigkeiten höherer Ordnung interpretiert werden können.

LITERATURVERZEICHNIS

American Educational Research Association, American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1-48. doi:10.18637/jss.v067.i01.

Bhavnani, S. K., Peck, F. A. & Reif, F. (2008). Strategy-based instruction: Lessons learned in teaching the effective and efficient use of computer applications. *ACM Transactions on Computer-Human Interaction (TOCHI), 15*, 1-47. doi: 10.1145/1352782.1352784

Binkley, M., Erstad, O., Herman, J., Raizen, R., Ripley, M. & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills (pp. 17 – 66)*. Dordrecht: Springer.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice, 5,* 7-74. doi: 10.1080/0969595980050102

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071. doi: 10.1037/0033-295X.111.4.1061

Calvani, A., Cartelli, A., Fini, A. & Ranieri, M. (2009). Models and instruments for assessing digital competence at school. *Journal of e-Learning and Knowledge Society, 4*(3). Retrieved from http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/288/270

Caissie, A. F., Vigneau, F. & Bors, D. A. (2009). What does the Mental Rotation Test measure? An analysis of item difficulty and item characteristics. *Open Psychology Journal, 2,* 94–102. doi: 10.2174/1874350100902010094

Carroll, J. M. & Rosson, M. B. (1987). Paradox of the active user. In J. M. Carroll (Ed*.), Interfacing thought: Cognitive aspects of human-computer interaction* (pp. 80–111). Cambridge, MA, US: The MIT Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22. http://dx.doi.org/10.1037/h0046743

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302. http://dx.doi.org/10.1037/h0040957

Day, S. B. & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*, 153–176. doi:10.1080/00461520.2012.696438

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F. & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28. doi: 10.18637/jss.v039.i12

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197. http://dx.doi.org/10.1037/0033-2909.93.1.179

Engelhardt, L., Goldhammer, F., Naumann, J. & Hartig, K. (2016, März). *Können induzierte Zielorientierungen smarte Bearbeitungswege in ICT-Umgebungen begünstigen?* Vortrag auf der 4. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Berlin, Deutschland.

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia, 13*(1), 93-107. Retrieved from http://www.openu.ac.il/Personal_sites/download/Digital-literacy2004-JEMH.pdf

Eshet-Alkalai, Y. E. & Amichai-Hamburger, Y. (2004). Experiments in digital literacy. *CyberPsychology & Behavior, 7*, 421-429. https://doi.org/10.1089/cpb.2004.7.421

Eshet-Alkalai, Y. & Chajut, E. (2010). You can teach old dogs new tricks: The factors that affect changes over time in digital literacy. *Journal of Information Technology Education: Research, 9*(1), 173-181. Retrieved from http://www.jite.org/documents/Vol9/JITEv9p173-181Eshet802.pdf

European Parliament and the Council (2006). Recommendation of the European Parliament and the Council of 18 December 2006 on key competences for lifelong

learning. *Official Journal of the European Union, L394*. Retrieved from http://www.alfa-trall.eu/wp-content/uploads/2012/01/EU2007-keyCompetencesL3-brochure.pdf

Ferrari, A., Punie, Y. & Redecker, C. (2012). Understanding digital competence in the 21st century: an analysis of current frameworks. *In 21st Century Learning for 21st Century Skills* (pp. 79-92). Springer Berlin Heidelberg. doi:10.1007/978-3-642-33263-0_7

Fraillon, J., Ainley, J., Schulz, W., Friedman, T. & Gebhardt, E. (2014). *Preparing for life in a digital age.* https://doi.org/10.1007/978-3-319-14222-7

Fraillon, J., Schulz, W. & Ainley, J. (2013). *International Computer and Information Literacy Study Assessment Framework.* Retrieved from http://research.acer.edu.au/ict_literacy/9

Goldhammer, F., Naumann, J. & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment, 29*, 263-275. https://doi.org/10.1027/1015-5759/a000153

Greiff, S., Krkovic, K. & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56*, 83-103. Retrieved from http://hdl.handle.net/10993/16434

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F. & Funke, J. (2013). Computer-based assessment of Complex Problem Solving: concept, implementation, and application. *Educational Technology Research and Development, 61*(3), 407-421.

Hahnel, C., Goldhammer, F., Naumann, J. & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior, 55*, 486–500. http://doi.org/10.1016/j.chb.2015.09.042

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63*, 43-49. doi: 10.1026/0033-3042/a000109

Hornke, L.F. & Habon, M.W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369–380. https://doi.org/10.1177/014662168601000405

International ICT Literacy Panel (2002). *Digital Transformation: A Framework for ICT Literacy*: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/ICTREPORT.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1-73. doi: 10.1111/jedm.12000

Kiefer, T., Robitzsch, A. & Wu, M. (2016). TAM: Test Analysis Modules. R package version 1.99-6. Retrieved from http://CRAN.R-project.org/package=TAM

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding What Was Measured* (ETS Research Report RR-01-25). Retrieved from https://www.ets.org/Media/Research/pdf/RR-01-25-Kirsch.pdf

Kretzschmar, A., Neubert, J. C., Wüstenberg, S. & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence, 54*, 55-69. https://doi.org/10.1016/j.intell.2015.11.004

Lorenzen, M. (2001). The land of confusion? High school students and their use of the World Wide Web for research. *Research Strategies, 18*, 151–163. doi:10.1016/S0734-3310(02)00074-5

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine, 178*, 107-114. https://doi.org/10.7205/MILMED-D-13-00213

Mulholland, T. M., Pellegrino, J. W. & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive psychology, 12*, 252-284. https://doi.org/10.1016/0010-0285(80)90011-0

National Higher Education ICT Initiative (2003). *Succeeding in the 21st Century: What higher education must do to address the gap in information and communication technology proficiencies*. Retrieved from http://www.ets.org/Media/Tests/Information_and_Communication_Technology_Literacy/ICTwhitepaperfinal.pdf.

OECD (2011). *PISA 2009 Results: Students on Line: Digital Technologies and Performance* (Volume VI). PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264112995-en

OECD (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. OECD Publishing. http://dx.doi.org/10.1787/9789264128859-en.

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en.

OECD (2014). *PISA 2012 results: What students know and can do – student performance in mathematics, reading and science* (volume I, revised edition, February 2014). PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). Considerations in Computer-Based Testing. In C. G. Parshall , J. A. Spray, L. Kalohn & T. Davey (Eds.), *Practical Considerations in Computer-Based Testing* (pp. 1–12). New York: Springer.

Perfetti, C. A., Rouet, J.-F. & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *the construction of mental representations during reading* (pp. 99–122). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Pfaff, Y. & Goldhammer, F. (2011, September). *Measuring individual differences in ICT literacy: Evaluating Online Information.* Paper presented at the 14th Biennial Conference of the European Association for Research on Learning and Instruction (EARLI), Exeter, United Kingdom.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology, 41,* 1-48. https://doi.org/10.1006/cogp.1999.0735

Richter, T., Isberner, M. B., Naumann, J. & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern. *Zeitschrift für Pädagogische Psychologie, 26*, 313-331. https://doi.org/10.1024/1010-0652/a000079

Richter, T., Naumann, J. & Horz, H. (2010). Das Inventar zur Computerbildung (revidierte Fassung*)*. *Zeitschrift für Pädagogische Psychologie*, *24*, 23-37. doi: 10.1024/1010-0652/a000002

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53*, 145-161. doi: 10.1002/asi.10017

Rölke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012* (pp. 344-353). Chesapeake, VA: AACE.

Scalise, K. & Gifford, B. R. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Teaching, Learning and Assessment, 4*(6). Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653/0.http://www.jtla.org

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12: Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12.* Göttingen: Hogrefe.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I. & Scherer, R. (2016). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review, 19*, 58-84. https://doi.org/10.1016/j.edurev.2016.05.002

Simon, H. A. & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist, 26*, 145-159. http://dx.doi.org/10.1037/h0030806

Singley, M. K. & Anderson, J. R. (1985). The transfer of text-editing skill. *International Journal of Man-Machine Studies, 22*, 403-423. doi:10.1016/S0020-7373(85)80047-X

Sireci, S. & Zenisky, A.L. (2006). Innovative items format in computer-based testing: In pursuit of construct representation. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 329-347). Hillsdale, NJ: Erlbaum.

Stadler, M., Niepel, C. & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior, 65*, 100-106. https://doi.org/10.1016/j.chb.2016.08.025

van Deursen, A. J. & Van Dijk, J. A. (2009). Using the Internet: Skill related problems in users' online behavior. *Interacting with Computers, 21*, 393-402. https://doi.org/10.1016/j.intcom.2009.06.005

van Deursen, A. J. & van Dijk, J. A. (2011). Internet skills and the digital divide. *New Media & Society, 13*, 893–911. doi:10.1177/1461444810386774

Volpert, W. (1982). The model of the hierarchical-sequential organization of action. In W. Hacker, W. Volpert and M. v. Cranach (eds.), *Cognitive and Motivational Aspects of Action* (pp. 35-51). Amsterdam: NHPC.

Watermann, R. & Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal of Psychological Assessment, 18*, 190-203. doi:10.1027//1015-5759.18.3.190

Wenzel, S.F.C., Engelhardt, L., Kuchta, K., Hartig, K., Naumann, J., Goldhammer, F.,… & Horz, H. (2015). *Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) in PISA (Schlussbericht des Projekts CavE-ICT-PISA)*. Abgerufen von der Website der Technischen Informationsbibliothek (TIB) der Leibniz Universität Hannover: http://edok01.tib.uni-hannover.de/edoks/e01fb15/838771823.pdf

Wenzel, S.F.C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F.,… & Horz, H. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) (CavE-ICT). In BMBF (Hrsg.). *Forschung in Ankopplung an Large-Scale Assessments* (pp. 161-180). Bonn, Berlin: BMBF.

Wilhelm, O., Schroeders, U. & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe.* Göttingen: Hogrefe.

Wilson, M., De Boeck, P. & Carstensen, C.H. (2008). Explanatory item response models: A brief introduction. In: Hartig, J., Klieme, E., Leutner, D. (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Hogrefe Publishing, Göttingen.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370-371.

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ANHANGVERZEICHNIS

Anhang A

Arbeit 1: Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Horz, H., Hartig, K., & Wenzel, S. F. C. (*submitted, International Journal of Testing*). A framework for the performance-based testing of ICT skills.

A framework for the performance-based testing of ICT skills

Lena Engelhardt[a], Johannes Naumann[b], Frank Goldhammer[a,c], Andreas Frey[d,e], Holger Horz[b], Katja Hartig[b], S. Franziska C. Wenzel[b]

[a]German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

[b]Goethe University Frankfurt am Main, Frankfurt, Germany

[c]Centre for International Student Assessment (ZIB), Germany

[d]Friedrich Schiller University Jena, Germany

[e]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

Correspondence concerning this article should be addressed to:

Lena Engelhardt, German Institute for International Educational Research (DIPF)

Phone: +49 69 24 708 754

E-mail: lengelhardt@dipf.de

Mailing addresses: j.naumann@em.uni-frankfurt.de; goldhammer@dipf.de; andreas.frey@uni-jena.de; horz@psych.uni-frankfurt.de; hartig@psych.uni-frankfurt.de; wenzel@psych.uni-frankfurt.de

Abstract

The exploration of competence, when solving tasks that require using information and communication technology (ICT skills) can be regarded as a broad domain. When ICT skills are to be measured, a thorough theoretical analysis of the domain is necessary. Besides ICT specific knowledge, problem solving and comprehension of text and graphics were specified as generic components of ICT skills. Theoretical assumptions about requirements when confronted with specific cognitive ICT tasks (i.e., access, manage, integrate, evaluate, create) were formulated within our framework. Based on these assumptions, performance-based ICT skills test items were developed and implemented following standardized construction rules. The framework's utility was investigated empirically by two means. First, item difficulties for 69 items were determined according to the Rasch model. For each cognitive ICT task the distribution of related item difficulties covered a comparable ability range suggesting a successful item construction. Second, process indicators were derived from log files representing test takers' interactions with the items. Examining these interactions confirmed that the items mostly elicited intended behavior. The results obtained underline the utility of the proposed framework for deriving items both from an item development and an implementation perspective.

Information and communication technology (ICT) skills are regarded as a key competence for lifelong learning (European Parliament and the Council, 2006), successful participation in the labor market (van Deursen & van Dijk, 2011), and for participation in political and societal debate. ICT literacy has thus even been labeled a "survival skill" (Eshet-Alkalai, 2004), and matters across the lifespan (Poynton, 2005).

When measuring ICT skills it is common to emphasize that, besides knowledge tests or questionnaires (e.g. Goldhammer, Gniewosz, & Zylka, 2016; Richter, Naumann, & Horz, 2010), a focus should be given to hands-on ICT skills – in that, they should be considered in terms of what individuals are actually capable of (International Computer and Information Literacy Study (ICILS); Jung & Carstens, 2015). ICT skills are the abilities required to successfully solve tasks that are bound to using ICT. ICT skills therefore comprise of skills and knowledge in operating technology. These skills are regularly described as also reliant on literacies that are generic and not specific to the ICT domain, such as reading (e.g. Calvani, Cartelli, Fini, & Ranieri, 2009; International ICT Literacy Panel, 2002; Fraillon & Ainley, 2010). In assessments such as ICILS, possible ICT tasks are presented as cognitive tasks that require, for instance, "accessing" or "managing" information. It is, however, *not* regularly considered how exactly these ICT tasks relate to those generic skills. This of course renders it impossible to have item development informed by theories of generic skills, like reading, and item difficulties predicted by theoretically derived item properties (cf. Embretson, 1983). Following these considerations, the first out of the three goals of this paper is to define what generic skills are used and to systematically relate established psychological theories of those skills to the context of ICT in an assessment framework, in order to derive task characteristics determining item difficulties for different information tasks.

If such a framework had previously been used to develop items, these items have to subsequently be translated into a suitable assessment format. ICT skills are best assessed in a

performance-based manner by means of computers, as this mode can be expected to allow for the best construct representation (Sireci & Zenisky, 2006). Highly authentic items based on software from everyday life cannot be easily integrated in larger assessments (Parshall, Spray, Kalohn, & Davey, 2002). Simulation environments, in contrast, can be rather effortful in development and designers are confronted with the question of which aspects have to be mapped in the simulation and which can be omitted (Mislevy, 2013, p.108). The balancing act between authenticity and an economically feasible implementation is crucial for the implementation process, starting with design principles of the environment up to the scoring of behavior. The second goal of this study is thus to elaborate on various implementation questions in order to propose how a simulated environment suitable to measure ICT skills in a performance-based way may look like.

The first two goals focus on conceptual issues, item development and implementation into a technology-based testing environment. The third goal is to examine empirically whether these conceptual issues were realized adequately by statistically analyzing item difficulties and aspects of test takers' behavior.

While all three goals contribute to the question of how to assess ICT skills, the first goal of providing a theoretical basis for a measureable construct is tied closest to the question of construct representation, a central issue for the validity of test score interpretation (Embretson, 1983; Kane, 2013). In achieving this goal, we aim at closing the gap in current research between operational assessment and underlying assumptions of tapped theoretical skills (cf. van Deursen & van Dijk, 2009, p.393).

First goal: Integration of the relevant theoretical aspects of ICT skills into one assessment framework

For most ICT tasks, being in command of skills for operating technology and ICT specific knowledge will not suffice to solve all ICT tasks successfully. Other generic skills

that are not unique to the ICT domain are also implicitly or explicitly assumed to be required for solving information problems in existing conceptualizations of ICT skills. These generic skills are named in other conceptualizations as "key competencies" such as reading, problem solving, numeracy, logical, inferential, and metacognitive skills (Calvani et al., 2009, p.186), as "cognitive skills", described as reading, numeracy, critical thinking, and problem solving (International ICT Literacy Panel, 2002, p.1), or as "intellectual capacity that is based on conventional literacies" (Fraillon & Ainley, 2010, p.8). While the requirement of, for instance, numerical skills depends strongly on task content, we assume that problem solving skills and skills needed for comprehension are important in every task, as information has to first be captured from the environment, and secondly, the information problem has to be solved by interacting with the environment. For this reason, comprehension and problem solving skills are considered an indispensable part of ICT skills together with ICT specific knowledge.

Components of ICT skills

We will now briefly introduce theories from the domains to describe what task characteristics may drive item difficulty.

Comprehension of text and graphics. Nearly all ICT tasks require the processing of symbolically represented information to some degree. Even in tasks that do not involve higher-order reading processes such as evaluating information for its truth (such as installing a software program on a computer), decoding, syntactic parsing, and semantic integration of words ("Do you want to proceed?") will be required. When using the Internet as an information resource, text comprehension comes into play as detailed in models such as the construction-integration (CI-) model of Kintsch (1998). This model describes the process of text comprehension as a cyclical interplay of bottom-up and top-down processes. Starting off with the physical representation of the text processes (such as letter and word recognition,

semantic parsing, and local coherence processes) will build a propositional representation of the content of a text (textbase model). This model is integrated in a top-down fashion with prior knowledge, resulting in a situation model. As the content of information can also be pictures or sounds, one can distinguish processing of visual-verbal (e.g. written language), visual-pictorial (e.g. iconic material), auditory-verbal (e.g. spoken language), and auditory-pictorial (e.g. sounds) information (cf. Integrated Model of Text and Picture Comprehension; Schnotz, 2005). The importance of also processing pictorial information is supported by the conceptualization of Eshet-Alkalai (2004), who proposed photo-visual literacy as one aspect of digital literacy. In ICT environments, such comprehension processes are important to identify menu items or folders, to discover editing functions, or to check search result pages.

Problem solving. In ICT environments, problem solving is required when performing tasks such as search inquiries or using an unfamiliar system. According to Simon and Newell (1971), problem solving takes place in a problem space, with nodes for different states of knowledge and different operators to achieve the next node. Problem solving comprises of "activities required to construct a problem space in the face of a new task environment", and "activities required to solve a particular problem" (Simon & Newell, 1971, p.154). Within the problem solving process, a problem solver decides which node to choose as a point for further investigation and which operator might be best to achieve a desired goal. This process of investigation can describe how users may perform a search on the web: entering a search term, comparing results, considering whether to go back to enter a new search term, or to navigate to a web page given in the search results. Brand-Gruwel, Wopereis, and Walraven (2009) propose a model for solving such an information problem using the internet (IPS-I model), which includes steps from defining the information problem up to organizing and finally presenting the relevant information. Task demands when solving problems might depend on the differences between the start and end state, the amount of nodes that need to be

visited or the available operators. Thus, it matters whether users only have to decide on one option out of a given set of options, or whether individuals have to identify these options first. The quantity and quality of such decisions that have to be made to complete a task will also determine the difficulty of the task.

ICT specific knowledge. According to Funke and Frensch (2007), domain-specific knowledge is highly important when solving problems. This opinion is supported by Kintsch (1998) who proposed different cognitive processes depending on the amount of available knowledge. Thus, ICT specific knowledge is a central aspect for item development, as knowledge can guide both cognitive processes - comprehension and problem solving. Individuals who possess the relevant knowledge to solve a task have, as a result of their knowledge, an increased chance to solve the task. Whether such ICT specific knowledge is required for task solution (for example, about spam markers or saving formats) should determine item difficulty.

Taken together, ICT skills are assumed to be based on these three skills. Tasks that require generic skills but no ICT specific knowledge are not of interest for this study, as test scores would not represent ICT specific skills. Tasks that require only ICT specific knowledge but no generic skills are also not the focus of the study as they are rather routine tasks and require no higher levels of ICT skills.

Task characteristics for developing ICT skills items

Various framework conceptualizations have already been suggested to organize the broad domain of ICT skills following different aims and purposes (see for an overview Ferrari, Punie, & Redecker, 2012; Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016). One very influential framework is the one proposed by the International ICT Literacy Panel (2002), where ICT literacy is seen as being in command of a set of ”critical components” (p. 3) in solving information tasks, labeled as “access(ing)”, “manage(ing)”, “integrate(ing)”,

"evaluate(ing)", and "create(ing)" information. The framework of the ICT Literacy Panel inspired other developments through recent years (cf. National Higher Education ICT Initiative, 2003; ICILS, Fraillon & Ainley, 2010) and overlaps with other ICT conceptualizations (Calvani et al., 2009; Eshet-Alkalai, 2004). Considering the central role of this framework, we chose to base our item development on these tasks. Going further, we aim at describing which ICT specific demands the *five cognitive ICT tasks* of accessing, managing, integrating, evaluating, and creating impose on the user. Based on the previously described skills, task characteristics are derived that should guide item development in order to systematically manipulate item difficulty.

Accessing. Accessing describes "knowing about and knowing how to collect and/or retrieve information" (International ICT Literacy Panel, 2002, p. 17). ICT specific demands often involve a multitude of ways to navigate ICT environments, which carries the potential problem that users may feel disorientated in the Internet (van Deursen, 2010). Unfamiliar navigational structures impede navigation (Chen, Pedersen, & Murphy, 2011) and the breadth, depth and topology of hypertext structure matters (DeStefano & LeFevre, 2007). A prototypical item that can be used to measure the individual skill in accessing information may be a task necessitating using a search engine in the database of a library with the objective to find a reasonable selection of books on a specific topic. An effective search process might be supported by task structures encouraging the use of more specific search terms, for instance, by utilizing various filtering options or more than one search field. If these structures were not available in the ICT environment provided, the problem solving process is less defined. As a result, less proficient users might conduct a disorganized search inquiry. Knowledge about search engines and experiences in specifying search terms then have to be applied to solve the task. As a consequence, accessing tasks should be harder if such knowledge is important for the task solution. Besides, good comprehension and problem

solving skills might be a prerequisite to solve these less-structured tasks and to learn from such situations.

Managing. Managing refers to "applying an existing organizational or classification scheme" (International ICT Literacy Panel, 2002, p. 17). ICT specific demands in managing information involve handling and dealing with more or less complex systems to accomplish a task. If the software is unfamiliar, users need to adapt their previous knowledge to the task (Calvani et al., 2009). The ease of transferring knowledge to the use of a new interface depends on the similarity of structures, surfaces and contexts (Day & Goldstone, 2012), or on whether general concepts exist (Singley & Anderson, 1985). A prototypical item that can be used to measure the individual skill to manage information would require for instance grouping e-mails into a folder structure or renaming a folder. If the functions to complete these tasks were not visible on the surface but have to be selected for, knowledge and experience about such programs can support the solution process, alongside problem solving skills and might drive item difficulties. Harder tasks may require involving knowledge, for instance about saving formats, or printing options.

Integrating. Integrating requires "interpreting and representing information. It involves summarizing, comparing and contrasting" (2002, p. 17). ICT specific demands can be described as the enormous amount of accessible information (Metzger, 2007) that requires integration of information obtained from different sources in a self-determined manner. The ease with which one forms coherence (cf. Kintsch, 1998) across information from different sources depends on the amount of information units (like websites, documents, or e-mails), the degree of comparative information, and the degree of inconsistencies between documents (Perfetti, Rouet, & Britt, 1999). The integration process is more complex if the information differs regarding the breadth (Bhavnani, Jacob, Nardine, & Peck, 2003) or contains conflicting information (Hämeen-Anttila et al., 2014). A prototypical item would be to decide

on a specific language course by comparing the websites of different courses regarding several criteria, for instance prices, dates, and reviews of former participants. Easy integration tasks would contain only a few information units, only a little or no contradicting information, and require rather common ICT specific knowledge. Tasks are harder if this integration process is more complex, and ICT specific knowledge has to be included. Such ICT knowledge can be for instance source information, or knowledge about the typical structure of websites.

Evaluating. Evaluating information involves "making judgments about the quality, relevance, usefulness, or efficiency of information" (2002, p. 17). To deal with the increasing amount of information on the Internet (Edmunds & Morris, 2000), users have to evaluate the value of incoming information (Whittaker & Snider, 1996). ICT specific demands are, for instance, evaluation of the trustworthiness of information (Lorenzen, 2001), as publishing on the Internet is not necessarily bound by peer-reviewing processes or editorial control. Based on a model of task-based relevance assessment and content extraction (TRACE model; Rouet, 2006), the information provided has to be combined with prior knowledge and evaluated. Evaluating information might depend on the ease of identifying relevant criteria. Criteria can be truth, guidance, accessibility, scarcity, and weight of information (Simpson & Prusak, 1995), but also structural (e.g. domain names) and message (e.g. objectivity) features (Pfaff & Goldhammer, 2011), or cues for quality, such as title, and authority (Rieh, 2002). A prototypical item that can be used to measure the individual skill in evaluating information could be to judge a set of e-mails in the inbox in regards to their relevance for another person. Besides capturing the content of e-mails, users may also have to include knowledge about spam markers and available information about the senders of the e-mails. Thus, the item difficulty is expected to depend on the availability of relevant information and the kind of knowledge (e.g. common vs. expert knowledge) that is required to solve the task.

Creating. Creating describes "generating information by adapting, applying, designing, inventing, or authoring information" (2002, p. 17). Presenting information adequately out of countless options for editing poses special demands on the user. Software for designing, painting, or scientific plotting has substantially enlarged the possibilities to create and transform knowledge into graphical material, compared to times where such tasks had to be carried out without the help of computers. Nevertheless, using this kind of software might require special intellectual skills (Horz, Winter, & Fries, 2009; Cox, Vasconcelos, & Holdridge, 2010). Bulletin boards, blogs and e-mails require a different kind of writing (McVey, 2008) compared to traditional off-line writing, and users are no longer solely recipients but also producers of user-generated content (van Dijck, 2009). Creating might be more difficult in the case of indefinite task instruction (see Cognitive Process Theory of Writing; Flower & Hayes, 1981), because then, users need to draw on their own knowledge (e.g. about standards and norms or editing functions). A prototypical item that can be used to measure the individual skill in creating information would be to adapt the text of an e-mail in order for it to fit a specific recipient. In harder items, users will not only need to adapt settings based on knowledge but will also have to realize the need for adaptation on their own. *Whether these derived task characteristics are indeed suitable to develop easy and hard ICT skills items will be evaluated within the third goal.*

Second goal: item implementation

Three implementation questions will now be addressed. These concern the advantages of simulation-based environments (1), and how to implement items (2) as well as suitable response formats (3) within such environments. It is obvious that ICT skills can be measured directly, if the assessment uses the same technology as the one needed, and the skills are applied in everyday life situations. In this case, the cognitive processes underlying specific ICT skills are elicited in the test situation. As a variety of operating systems, their versions,

programs, and program versions exist, we suggest that a simulation environment in an ICT context should be, at first, abstracted from currently existing software. We regard this as advantageous for two reasons. The first is that the behavior of persons who typically use different operating systems can be compared without potential influences stemming from the simple fact that a person is unfamiliar or familiar with an operation system or a special kind of software. Second, using abstracted software will increase the potential to use a simulation environment for a longer time period, whilst using common operating systems would mean they become outdated as soon as newer versions are released. Moreover, we assume that adapting to new systems, which a simulation environment would be, is an integral part of ICT skills and does not conflict with the construct interpretation.

For implementation, however, many decisions regarding design choices and the degree of complexity have to be made (cf. Parshall et al., 2002, p.10). A second and central question is how authentic items have to be to capture the cognitive processes test developers are interested in, and at the same time, avoid the same needless expenses that do not improve construct representation. As a "higher fidelity of real-world situation does not necessary make for better assessment" (Mislevy, 2013, p.108), we suggest restricting the environment, in that just as many options can be implemented in a simulation, as they would under realistic settings, so that the same cognitive processes need to be used to solve an item. This means that not only the solution process but also options to solve the item incorrectly should be implemented while parts of the reality that are irrelevant to item solution will not be implemented. For instance, a browser search would not be available when the task is to evaluate e-mails. Test takers will not realize the limitations of the environment as long as they work on the item as suggested by the instructions. The theoretical differentiation into cognitive tasks such as accessing or evaluating information allows asking only for a specific goal in items (e.g. evaluating the trustworthiness of e-mails) that may contain several, but

limited, reactions. This also allows that test takers decide independently to move to the next item, as test takers should not get "lost" in the items if reactions are limited to the aspects of interest. An explicit instruction and previously conducted thinking aloud try-outs and calibration studies can help to identify problematic aspects of items or items being problematic in general.

Implementation also involves considering suitable response formats for scoring. At first, cognitive processes are closest to realistic settings if behavior that is typically shown in ICT contexts is required, for instance to forward an e-mail after evaluating its relevance for another person, instead of only marking the relevant e-mails in a multiple choice format. Using only items that can be automatically and immediately scored allows for instant test score estimation and reporting as well as computerized adaptive testing possible. As a pre-requisite of instant scoring, item instructions have to define the task so clearly that the test takers show only the intended correct or incorrect behavior. Or, possible items responses have to be restricted by the environment, for instance by using selective rather than constructive answering formats (cf. Scalise & Gifford, 2006). Selective but authentic response formats in ICT items would be behaviors such as saving or moving of the targeted documents, selecting a specific saving format out of given options, or deciding whether or which program should be downloaded. Constructive formats would require, for instance, performing a search inquiry or naming a document and can only be scored automatically if the goal is clearly defined in instruction. *Whether these implementation principles are suitable to implement items where test takers reason about the task as intended will be answered in the following within the third goal.*

Third goal: application and evaluation of the framework

The following section aims at addressing the appropriateness of the conceptual ideas presented above.

Empirical research question

It was empirically examined in how far the suggested theoretical framework and the derived conceptual basis for item development and implementation were useful to develop a measurement instrument.

Empirical research question 1 - item difficulties. As task characteristics driving item difficulties were derived from a theoretical basis, the first research question asks: Is the proposed framework suitable to develop items that address each of the five cognitive tasks across a wide ability range? For instance, if items showed only a high difficulty for a particular cognitive task, the framework would not be helpful when attempting to manipulate item difficulty systematically.

Empirical research question 2 - test-taking process. As behavior constitutes a link between the actually performed cognitive processes and the item scores, the second research question asks: Did the test takers behave as intended in the implemented items? For instance, if many test-takers showed less interaction than expected for following the instruction appropriately, this could point to a malfunctioning item.

Method

Items. Seven item developers commented reciprocally on item drafts to ensure that only items fitting into the framework were given to item implementation. They also specified expected item difficulties and reasons for these. Table 1 contains prototypical items for the five cognitive tasks, the ICT specific skills they require, and empirical item difficulties.

--- Table 1 about here---

For implementation, the authoring tool CBA ItemBuilder was used (Rölke, 2012), which allowed item developers to simulate environments (e.g. browser, e-mail inbox, text processing and others) and to implement scoring rules themselves. Cognitive laboratories ensured the usability of the testing environment and the clarity of all elements of the

constructed items (for a fully implemented item, see: Engelhardt, Goldhammer, Naumann, & Frey, 2017; and for further information about the study, see: Wenzel et al., 2016).

Sample and data collection. The sample consisted of $N = 773$ students from 34 schools in Germany. Students were on average 15.29 years old ($SD = 0.66$) and about half of them were male (male: 51%, female: 46%, not specified: 3%). Test takers received randomized subsets of about 33 items, as all 70 items would have taken too much time for one test taker to answer. Test takers worked on average $M = 105.59$ seconds ($SD = 40.04$) on an item. The items were scored dichotomously (correct/incorrect) immediately after a response was given to a task and indicators of response behavior (i.e., the number of interactions) were automatically extracted from the log data. To discover discrepancies in the theoretically assumed solution behavior, the minimum number of interactions with the environment for a correct item solution was computed for each of the items.

Data analyses. A one-dimensional Rasch model was fitted using the R package TAM (Kiefer, Robitzsch, & Wu, 2016; R Core Team, 2014) with the mean of the ability distribution set to 0 and the slope of the item characteristic curves set to 1 (so the ability variance was freely estimated). One item was excluded because of an insufficient item-fit (Outfit: 2.18; cf. de Ayala, 2013); for all other items the item-fit was acceptable (range of Infit: .87-1.11; range of Outfit: 0.67-1.25). The expected a priori (EAP) reliability for the remaining 69 items was 0.71. For the first empirical research question, Levene's test for homogeneity of variance and an ANOVA were conducted for items from the five cognitive tasks, with item difficulties as dependent and the five cognitive tasks as independent variables. For the second empirical research question, the range of the number of interactions was compared with the minimum number of interactions for a correct item solution per item. Specifically, the range from the 1st to the 3rd quartile was used to capture how a considerable amount of test takers behaved.

Results and discussion

Empirical evidence 1 - item difficulties. The 69 items had an average difficulty of M = 0.38 (SD = 1.56) and were distributed across a wide ability range (Min = -2.84; Max = 4.27). Levene's test ($F(4, 64) = 0.13$, $p = .971$) and the ANOVA ($F(4, 64) = 1.12$, $p = .356$) indicated that variances for items from different cognitive tasks were homogeneous and did not differ in their average item difficulties, supporting a successful item construction (cf. Figure 1).

-- Figure 1 about here--

Empirical evidence 2 - test-taking process. Figure 2 shows for each of the 69 items a comparison of the number of interactions in-line with the theoretically assumed solution behavior (dashed line) with the observed number of user interactions.

-- Figure 2 about here --

For three items (A-C), more than 25 % of test takers interacted less with the items than it would have been necessary for a correct solution, indicating that a considerable amount of test takers did not perform the cognitive processes as expected. However, in-depth analyses for items A and B revealed that the expected incorrect solutions required fewer interactions than the correct solution. As the item difficulties of item A (2.37) and item B (1.68) were high, many test taskers might have chosen the incorrect way, which can explain the small number of interactions and does not necessarily question the quality of the items. Instead, for item C (item difficulty: 0.47), the expected correct and incorrect solution required the same amount of interactions. Thus, choosing the incorrect solution is no explanation for the small number of interactions. Analyzing logdata revealed that many test takers visited only one of the two required websites to solve the item. Thus, they did not perform the cognitive processes as expected, which would have been to compare the information on both websites. Modifying the instruction in such a way that the test takers are required to visit both websites may help to improve item functioning. Two further items are conspicuous: Most test takers

required for items D and E far more interactions on average than the minimum amount, but differed also strongly from each other (cf. large range) than in the other items. This can be technically explained due to required scrolling in the items and is not a problem for the functioning of the item. Still, processing times were rather high and diverse (D: $M = 241$, $SD = 130$; E: $M = 154$, $SD = 115$), which can be problematic as time is precious in assessments. After thorough inspection we decided to revise those items and to reduce their complexity by removing information units, like the number of e-mails test takers had to deal with.

General discussion

The goal of this study was to propose a theoretical framework enabling a theory-driven item development and implementation. The framework allowed us to evenly capture the targeted construct of ICT skills as indicated by comparable mean levels and ranges of item difficulties for all five cognitive ICT tasks. Furthermore, the test-taking behavior elicited by the items was as intended for nearly every item. Only a few items (4.3%) required adaptation, in terms of further specifying the instructions or by decreasing the amount of presented information, such as e-mails.

Our framework extends previous work in that the five cognitive ICT tasks that have been proposed before were not only defined operationally but anchored into established theories and described in terms of what drives item difficulties. Empirical support that the ICT skills test is eliciting the intended cognitive processes, based on the suggested theoretical framework, is already available. In an experimental validation study (Engelhardt et al., 2017) it was examined whether the described task characteristics actually determined item difficulties as intended. Nevertheless, further research is needed to empirically investigate the extent and kind of relationship between the construct of ICT skills and the constructs of problem solving, comprehension of text and graphics and technological knowledge.

As computer-based performance assessments become more the rule rather than the exception, we suggest investigating whether test takers perform the assumed cognitive processes by considering process indicators like user interactions. Also further indicators are plausible to investigate the functioning of items. These can be processing times (e.g. excluding rapid guessing behavior) or to gather more information about the behavior in the items based on logdata. For instance, analyzing whether sub-goals were reached or which kinds of mistakes were made.

In the present study we demonstrated that so-called new domains can be related back to theories about well-established constructs. Strategies for implementation completed the foundation of the item development process. We assume that the presented approach of item development and implementation is not only useful for the assessment of ICT skills, but also for other contemporary constructs, such as 21st century skills, being assessed by computer-based simulations.

References

Bhavnani, S. K., Jacob, R. T., Nardine, J., & Peck, F. A. (2003). Exploring the distribution of online healthcare information. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems* (pp. 816-817). ACM.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education, 53*, 1207-1217. https://doi.org/10.1016/j.compedu.2009.06.004

Calvani, A., Cartelli, A., Fini, A., & Ranieri, M. (2009). Models and instruments for assessing digital competence at school. *Journal of e-Learning and Knowledge Society-English Version*, *4*(3). Retrieved from http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/288/270

Chen, C.-Y., Pedersen, S., & Murphy, K. L. (2011). Learners' perceived information overload in online learning via computer-mediated communication. *Research in Learning Technology, 19*, 101–116. doi:10.1080/21567069.2011.586678

Cox, A. M., Vasconcelos, A. C., & Holdridge, P. (2010). Diversifying assessment through multimedia creation in a non-technical module: Reflections on the MAIK project. *Assessment & Evaluation in Higher Education, 35*, 831–846. doi:10.1080/02602930903125249

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*, 153–176. doi:10.1080/00461520.2012.696438

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

DeStefano, D., & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior, 23*, 1616–1641. doi:10.1016/j.chb.2005.08.012

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management, 20*, 17–28. doi:10.1016/S0268-4012(99)00051-1

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197. http://dx.doi.org/10.1037/0033-2909.93.1.179

Engelhardt, L., Goldhammer, F., Naumann, J. & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. *Computers in Human Behavior, 76*, 683-692.

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia*, *13*, 93–107. Retrieved from http://www.openu.ac.il/Personal_sites/download/Digital-literacy2004-JEMH.pdf

European Parliament and the Council (2006). Recommendation of the European Parliament and the Council of 18 December 2006 on key competences for lifelong learning. *Official Journal of the European Union, L394*. Retrieved from http://www.alfa-trall.eu/wp-content/uploads/2012/01/EU2007-keyCompetencesL3-brochure.pdf

Ferrari, A., Punie, Y., & Redecker, C. (2012). Understanding digital competence in the 21st century: an analysis of current frameworks. In *21st Century Learning for 21st Century Skills* (pp. 79–92). Springer Berlin Heidelberg. doi:10.1007/978-3-642-33263-0_7

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32,* 365–387. doi: 10.2307/356600.

Fraillon, J., & Ainley, J. (2010). *The IEA international study of computer and information literacy (ICILS)*. Retrieved from

http://www.researchgate.net/profile/John_Ainley/publication/268297993_The_IEA_International_Study_of_Computer_and_Information_Literacy_%28ICILS%29/links/54eba4330cf2082851be49a9.pdf

Funke, J., & Frensch, P. A. (2007). Complex Problem Solving: The European Perspective - 10 Years After. In D. H. Jonassen (Ed.), *Learning to Solve Complex Scientific Problems* (pp. 25–47). New York, NY, US: Lawrence Erlbaum Associates.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment*, *29*, 263–275. https://doi.org/10.1027/1015-5759/a000153

Goldhammer, F., Gniewosz, G., & Zylka, J. (2016). ICT Engagement in learning environments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 331–351). Dordrecht: Springer International Publishing.

Hämeen-Anttila, K., Nordeng, H., Kokki, E., Jyrkkä, J., Lupattelli, A., Vainio, K., & Enlund, H. (2014). Multiple information sources and consequences of conflicting information about medicine use during pregnancy: a multinational Internet-based survey. *Journal of Medical Internet Research*, *16*. doi:10.2196/jmir.2939

Horz, H., Winter, C., & Fries, S. (2009). Differential benefits of situated instructional prompts. *Computers in Human Behavior*, *25*, 818–828. doi:10.1016/j.chb.2008.07.001

International ICT Literacy Panel (2002). *Digital Transformation: A Framework for ICT Literacy*: ETSRetrieved from http://www.ets.org/Media/Research/pdf/ICTREPORT.pdf

Jung, M., & Carstens, R. (Eds.) (2015). *ICILS 2013 user guide for the international database.* Amsterdam: IEA.

http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICILS_2013_IDB_user_guide.pdf. Last accessed 23rd August 2017.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73. doi: 10.1111/jedm.12000

Kiefer, T., Robitzsch, A., & Wu, M. (2016). TAM: Test Analysis Modules. R package version 1.99-6. Retrieved from http://CRAN.R-project.org/package=TAM Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Lorenzen, M. (2001). The land of confusion? High school students and their use of the World Wide Web for research. *Research Strategies, 18*, 151–163. doi:10.1016/S0734-3310(02)00074-5

McVey, D. (2008). Why all writing is creative writing. *Innovations in Education and Teaching International, 45*, 289–294. doi:10.1080/14703290802176204

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*, 2078–2091. doi: 10.1002/asi.20672

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine, 178*, 107–114. https://doi.org/10.7205/MILMED-D-13-00213

National Higher Education ICT Initiative (2003). *Succeeding in the 21st Century: What higher education must do to address the gap in information and communication technology proficiencies.* Retrieved from http://www.ets.org/Media/Tests/Information_and_Communication_Technology_Literacy/ICTwhitepaperfinal.pdf.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Considerations in Computer-Based Testing. In C. G. Parshall , J. A. Spray, L. Kalohn, & T. Davey (Eds.), *Practical Considerations in Computer-Based Testing* (pp. 1–12). New York: Springer.

Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Pfaff, Y., & Goldhammer, F. (2011, September). *Measuring individual differences in ICT literacy: Evaluating Online Information.* Paper presented at the 14th Biennial Conference of the European Association for Research on Learning and Instruction (EARLI), Exeter, United Kingdom.

Poynton, T.A. (2005). Computer literacy across the lifespan: A review with implications for educators. *Computers in Human Behavior, 21*, 861–872.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Richter, T., Naumann, J. & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R). *Zeitschrift für Pädagogische Psychologie, 24*, 23–37. doi: 10.1024/1010-0652/a000002

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53*, 145–161. doi: 10.1002/asi.10017

Rölke, H. (2012). The item builder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of World Conference on E-*

Learning in Corporate, Government, Healthcare, and Higher Education (pp. 344–353). Chesapeake, VA: AACE.

Rouet, J.-F. (2006). *The skills of document use: From text comprehension to web-based learning.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Scalise, K., & Gifford, B. R. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Teaching, Learning and Assessment, 4*(6).Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653/0.

Schnotz, W. (2005). An Integrated Model of Text and Picture Comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–69). New York, NY, US: Cambridge University Press.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review, 19*, 58–84. https://doi.org/10.1016/j.edurev.2016.05.002

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist, 26*, 145–159. http://dx.doi.org/10.1037/h0030806

Simpson, C. W., & Prusak, L. (1995). Troubles with information overload—Moving from quantity to quality in information provision. *International Journal of Information Management*, *15*, 413–425. doi:10.1016/0268-4012(95)00045-9

Singley, M. K., & Anderson, J. R. (1985). The transfer of text-editing skill. *International Journal of Man-Machine Studies, 22*, 403–423. doi:10.1016/S0020-7373(85)80047-X

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.),

Handbook of Test Development (pp. 329-347). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

van Deursen, A. J. (2010). *Internet skills: vital assets in an information society*. University of Twente, Netherlands. http://dx.doi.org/10.3990/1.9789036530866

van Deursen, A. J., & van Dijk, J. A. (2009). Using the Internet: Skill related problems in users' online behavior. *Interacting with Computers, 21*, 393–402. doi:10.1016/j.intcom.2009.06.005

van Deursen, A., & van Dijk, J. (2011). Internet skills and the digital divide. *New Media & Society, 13*, 893–911. doi: 10.1177/1461444810386774

van Dijck, J. (2009). Users like you? Theorizing agency in user-generated content. *Media, Culture & Society, 31*, 41–58. doi: 10.1177/0163443708098245

Wenzel, S.F.C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F., Naumann, J., & Horz, H. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) (CavE-ICT) [Computer-based, adaptive and behavior-related assessment of information and communication-related competencies (ICT skills)]. In BMBF (Hrsg*.). Forschung in Ankopplung an Large-Scale Assessments* (pp. 161-180). Bonn, Berlin: BMBF.

Whittaker, S., & Sidner, C. (1996, April). E-mail overload: exploring personal information management of e-mail. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 276–283). ACM.

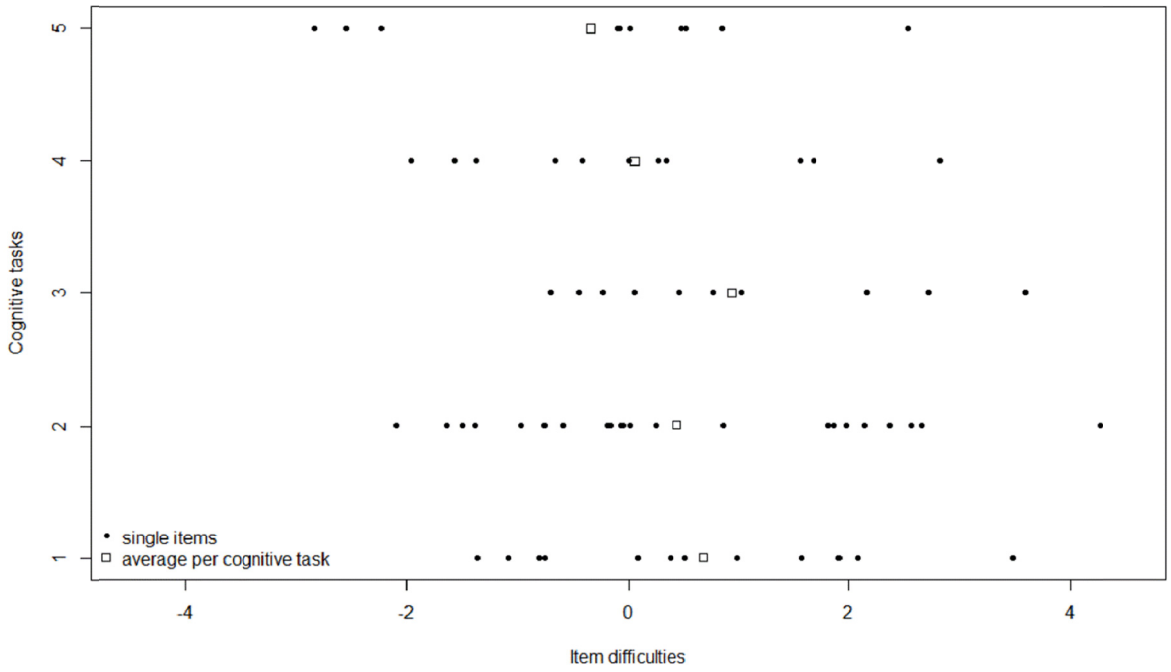


*Figure 1.*Empricial item difficulties of items assessing the five cognitive tasks, access (1), manage (2), integrate (3), evaluate (4), and create (5).

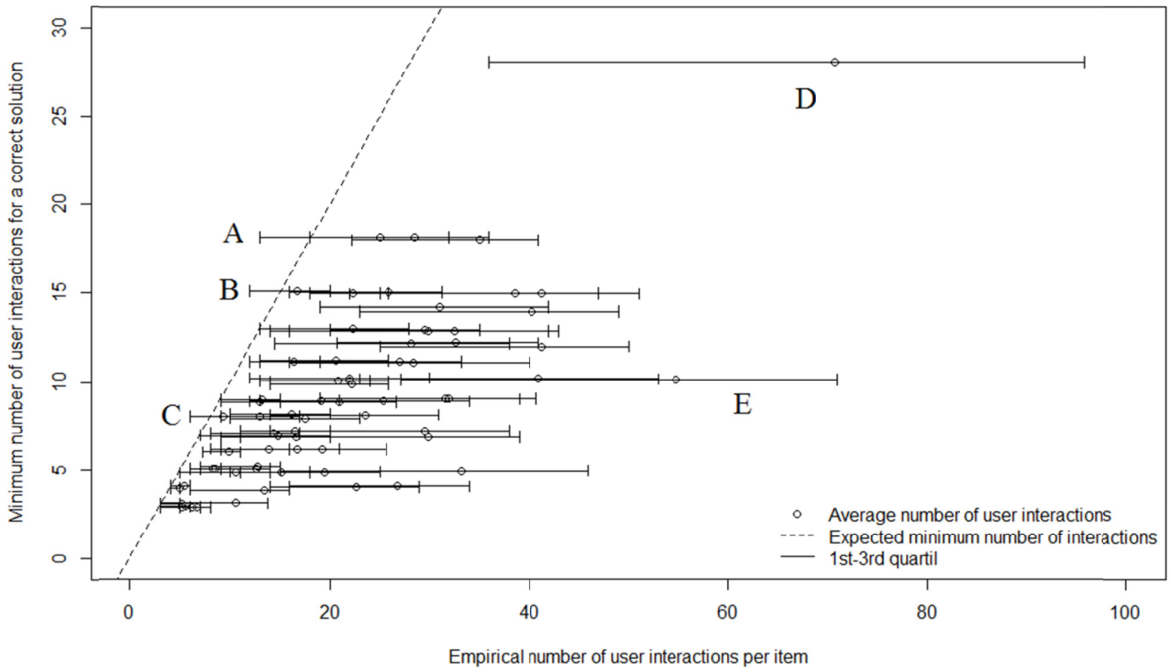


Figure 2. Empirical user interactions compared to the theoretical minimum number of user interactions needed for a correct item solution.

Table 1

Prototypical Items.

Cognitive ICT tasks	Item description	ICT specific skill for 5 cognitive ICT tasks	Characteristics increasing (+) or decreasing (-) item difficulty	Assumed difficulties	Empirical difficulties
Access	The student needs to search in a libraries' database for books for a specified topic.	The student is able to refine his inquiry to find the targeted books by using more than one key word.	optional literature has to be identified first, no preselection was presented (+) search term has to be entered without preselected options (+)	advanced	3.47
Manage	The student has to rename and to move fitting e-mails into a folder in his e-mail inbox.	The student is able to identify the correct folder and figures out how to rename and move e-mails in the e-mail inbox.	rename function easy to find/ familiar structures (-) folder and e-mails are easy to identify (-)	easy	-1.64
Integrate	The student needs to decide for one out of two language courses based on several criteria.	The student is able to identify and compare the relevant aspects.	decisive information is scattered across two web sites(+) only one decision based on two preselected course options (-)	medium	0.47
Evaluate	The student has to decide for 5 e-mails whether they have to be forwarded to a new colleague and forwards them if necessary.	The student is able to identify markers of spam to decide correctly not to forward the hoax e-mail.	knowledge about hoax e-mails is required (+) spam markers hard to identify, because hoax e-mail was sent by a colleague (+) many options: decision for 5 e-mails required (+)	advanced	2.81
Create	The student needs to reply in an adequate manner to an e-mail invitation by changing available text components and icons.	The student adapts text and icons in order that they are appropriate for the situation.	knowledge about norms and values in official contexts is required (+) two decisions (about text and icon) based on preselected options (-)	medium	0.09

Anhang B

Arbeit 2: Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Wenzel, S. F. C., Hartig, K., & Horz, H. (*submitted, European Journal of Psychological Assessment*). Convergent evidence for validity of a performance-based ICT skills test.

Convergent evidence for validity of a performance-based ICT skills test

Lena Engelhardt[a], Johannes Naumann[b], Frank Goldhammer[a,c], Andreas Frey[d,e], S. Franziska C. Wenzel[b], Katja Hartig[b], Holger Horz[b]

[a]German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

[b]Goethe University Frankfurt am Main, Germany

[c]Centre for International Student Assessment (ZIB), Germany

[d]Friedrich Schiller University Jena, Germany

[e]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

Correspondence concerning this article should be addressed to:

Lena Engelhardt, German Institute for International Educational Research (DIPF)

Phone: +49 69 24 708 754

Email: lengelhardt@dipf.de

Abstract

The goal of this study was to investigate convergent sources of validity evidence supporting the construct interpretation of a test score from a simulation-based ICT skills test. The construct definition understands ICT skills as reliant on ICT specific knowledge and two generic skills, comprehension and problem solving skills. Based on this, three validity arguments were formulated and tested. Applying the classical nomothetic span approach, all three predictor variables explained task success positively across all ICT skills items (1). As ICT tasks can vary in the extent to which they require construct-related knowledge and skills and in the way related items are designed and implemented, varying effects of construct-related predictor variables across items were expected. A task-based analysis approach revealed that the item-level effects of the three predictor variables were in line with the targeted construct interpretation for most items (2). Finally, item characteristics could explain the random effect of problem solving skills significantly; however, this was not possible for comprehension skills (3). Taken together, the obtained results support the validity of the targeted construct interpretation.

Skills to handle information and communication technologies (ICTs) are described as 21st century skills (Binkley, Erstad, Herman, Raizen, Ripley, & Rumble, 2012), as survival skills (Eshet-Alkalai, 2004), and as key competencies for lifelong learning (European Parliament and the Council, 2006). Due to their importance, they are targeted in assessments (Ferrari, Punie, & Redecker, 2012). Variables used as a source for convergent validity evidence of ICT skills are mostly demographic in nature, such as gender, socio-economic status, or self-reports concerning the use of ICT and self-efficacy (Siddiq, Hatlevik, Olsen, Throndse, & Scherer, 2016, p.33). We suggest that one reason for this is because theoretical assumptions about involved skills are somewhat vaguely described, thus a clearly defined construct definition that includes the type of skills required and also their interplay is missing. Many conceptualizations consider ICT skills as a mixture of technical proficiency and other abilities that are not exclusive to ICT contexts, but differ in how they describe and label these additional skills. Additional skills are labeled as reasoning, metacognitive skills, critical thinking, reading, problem solving, numerical skills (Calvani, Cartelli, Fini, & Ranieri, 2009, p.186; International ICT Literacy Panel, 2002, p.1), or more abstractly as "conventional literacies" (Fraillon & Ainley, 2010, p.8). In light of these issues, it is not surprising that such variables are not conventionally considered as convergent sources of validity evidence. As the validity of the construct interpretation is targeted in this study, the first goal of this study is to specify those skills and also their interplay to formulate testable validity arguments.

A second challenge in considering convergent sources of validity evidence for ICT skills test scores is the broadness of possible ICT tasks that are represented in the items. Nearly every information task can be performed using ICTs. ICT tasks might require evaluating the trustworthiness of information or managing a numerical table. As a consequence, required skills, such as reading, might vary in relevance to different tasks as they might require to a greater or lesser extent reading processes. When investigating the correlation between test

scores and involved skills, the correlation might vary for different items. Variations alone do not necessarily challenge the validity of test score interpretation. However, if a few items required for instance no ICT-specific but only generic skills, one could question whether test scores from these items represent ICT skills at all. The second goal of this study is to apply a task-based approach to investigate whether construct interpretation is also valid on an item level.

In this study, we focus on a simulation-based ICT skills test that should assess 15 year old students' ability to handle everyday ICT tasks. The goal of this study is to provide convergent sources of validity evidence of test scores for construct interpretation, based on the understanding of validity as defined by AERA, APA and NCME (2014), where not the test itself but the targeted interpretation is validated.

ICT Skills

We start off with describing the targeted construct interpretation. Then, the relevant skills and their interplay are specified, which build the basis for formulating testable hypotheses.

Targeted Construct Interpretation of the Test Scores

Performing tasks in ICT environments requires skills on various levels. While every ICT task requires at least interacting with the environment via mouse clicks or typing, some tasks also require making decisions based on complex considerations. An ICT task that requires basic skills (cf. Basic Computer Skills; Goldhammer, Naumann, & Kessel, 2013) in operating technology could be, for instance, to forward a predefined email. A task that requires higher order ICT skills would also contain, besides the forwarding process of the email, the decision of whether the email should be forwarded or not. To make such a decision, spam markers or

credibility criteria of the email have to be detected and taken into account to make the decision. The ability to perform this decision-making should be represented in the test scores. ICT tasks that include these types of decisions might require choosing relevant books in a library database, deciding between two language courses based on the content of their websites, or adapting an email so that it fits to social norms. Higher order ICT skills that are required to make these decisions should be reflected in the test scores of the performance-based ICT skills test.

Assumed Cognitive Processes in Solving ICT Tasks

We assume that ICT specific skills, such as ICT specific knowledge, play a central but not exclusive role in solving ICT tasks. Regarding the understanding of the information presented in the environment, we assume that cognitive processes consist of comprehending textual (e.g. words) and graphical (e.g. pictures) information (Schnotz, 2005). Based on Kintsch's (1998) construction-integration (CI-) model, text comprehension starts with lower processes to process letters and words and then requires to build a propositional representation of the text's contents (textbase model). In further processes, prior knowledge is then integrated to construct a situation model. For instance, when processing textual material, the reading load of a task depends on the characteristics and determines the extent to which reading comprehension processes occur. To handle the environment, to perform and organize different steps to reach a defined goal, we assume that the cognitive processes are similar to the processes in problem solving (Simon & Newell, 1971). Simon and Newell describe problem solving as taking place in a problem space where problem solvers have to figure out their way by selecting different operators to reach a certain goal state. Although problem solving occurs internally, behavior can serve as an indirect indicator for the cognitive processes that are performed (Mayer & Wittrock, 2006). The extent to which problem solving

processes occur in a task depends on its "intrinsic complexity", which can be described in terms of a minimum number of steps or diversity of behavior (OECD, 2012, p.50). We assume that ICT specific knowledge can help to guide comprehension and problem solving processes, such as seeking specific information in a database or figuring out how to forward an email in a new environment. Figure 1 describes these different skills that are assumed to be relevant for solving ICT tasks. The dark grey area describes how ICT skills are understood in this study. We focus on higher-order ICT skills, which are required to make correct decisions in ICT tasks and are based on generic comprehension and problem solving skills as well as on ICT specific knowledge.

--- Insert Figure 1 about here ---

Whether, and to what extent these generic skills are needed depends strongly on the characteristics of the task. Words and information units, for example, determine whether comprehension processes are evoked and the length and structure of a task, between its start and end state, determines whether problem solving processes are evoked. A task within spreadsheet or presentation software might contain less text and might thus require less comprehension skills than a task in browser environments. Searching a document on the computer in a complex and deep folder structure may lead, depending on the location of the document, to a solution path of different length. Longer solution paths should require more problem solving skills. ICT specific knowledge is required in every task, but might vary in difficulty. Knowledge about spam markers might be less well known and harder to apply compared to knowledge about the functioning of email inboxes, especially for younger people.

Tasks being primarily hard, because they require higher levels of comprehension or problem solving skills are not those ICT tasks that are targeted in this study. ICT tasks should

be focused that are rather hard because more advanced ICT knowledge has to be applied and integrated into task solution.

Convergent Sources of Evidence for the Construct Interpretation

In the following, convergent sources of evidence are defined. We start off by following the classical nomothetic span approach (Embretson, 1983) that assumes relations on test score level. In the standards of validity (AERA, APA, & NCME, 2014), such an approach is labeled as validity based on relations to other variables. Evidence should be provided that persons with higher test scores can be testified as possessing higher ICT skills in sense of the above described definition (cf. Figure 1). *The construct interpretation of the test score is supported, if unique positive effects on solving ICT skills items can be found for ICT specific knowledge, problem solving, and comprehension skills (Hypothesis 1).*

Even if the probability of success in ICT skills items was on average positively predicted by construct-related knowledge and skills, this might not necessarily be true for every single item due to item-specific composites of task characteristics. Task characteristics determine the type and extent of knowledge and skills that are required for task solution. The variation of the effect of a skill across items might depend on how the items were designed at the surface level (e.g., the amount and complexity of text presented to the test-taker). Varying effects do not question the item design as long as they are in line with the intended construct interpretation. According to Figure 1, ICT specific knowledge should influence the probability of success in any ICT skills item. Additionally, problem solving and/or comprehension skills should play a role. Extending the classical nomothetic span approach, the relations of construct-related knowledge and skill variables should also be investigated at item level to clarify whether the pattern of effects fits the intended construct interpretation. *The construct interpretation of test (and item) scores is supported, if positive relations to task*

success for all items can be found; either for ICT specific knowledge and one of the two generic variables (problem solving or comprehension sills) or for all three variables (Hypothesis 2).

Varying effects of construct-related knowledge and skills should be caused by varying item characteristics. In the construct representation approach (Embretson, 1983) item characteristics that should evoke the targeted processes for construct interpretation are quantified. A relation of indicators to item difficulties then supports the targeted construct interpretation. Item characteristics evoking reading comprehension and problem solving processes can also be found in ICT skills items. The reading load of an item can be assumed to indicate required reading comprehension processes, and the number of steps in a solution process can indicate required problem solving processes. For ICT specific knowledge in contrast it is less easy to identify quantifiable task characteristics (Engelhardt, Goldhammer, Naumann, & Frey, 2017). These task characteristics for comprehension and problem solving skills are alone not sufficient to support construct interpretation, as they only capture generic demands, while ICT-specific demands are not represented. However, they are important as they can support test score interpretation if they co-vary with the varying effects of the generic skills. As a result, they would support that reading and problem solving processes occur in the items. *The construct interpretation of test scores is supported, if the strengths of the positive effects of comprehension and problem solving skills depend on quantifiable item characteristics that determine how strongly these cognitive processes are evoked (Hypothesis 3).*

Method

Sample and Procedures

The sample consisted of $N = 269$ 15 year old German students ($M = 15.29$, $SD = .68$) who were roughly split equally male (52%) and female (46%; rest not specified). They were from 34 German schools from two federal states in Germany. The assessment consisted of two parts, where each part took about one hour. Before the test-takers started with the test, they received a tutorial to become familiar with the simulated environment. Then, students were assigned randomly to four different booklets. Each student first worked on around 33 of the 70 items. In the second part of the assessment, test-takers received questions to assess ICT specific knowledge, the reading comprehension test, and finally the problem solving test. Data from 256 students who finished all tests was used for analyses.

Measures of Person Variables

ICT skills items were developed following the conceptualization of the International ICT Literacy Panel (2002), who distinguished occurring ICT tasks into accessing, managing, integrating, evaluating and creating information. They were implemented in a simulation environment by means of the CBA ItemBuilder (Rölke, 2012). Applications like browsers, e-mail inboxes or file managers were simulated. The developed performance items were scored dichotomously immediately after item response. After excluding six items due to item fit and differential item functioning (Wenzel et al., 2016), 64 items were selected for the test.

Problem solving was assessed with seven items from the scale “Complex Problem Solving” (MicroDYN; Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013). For each item, scores for knowledge acquisition (EAP Reliability: .77) and knowledge application (EAP Reliability: .75) were extracted and fitted in a two-dimensional two-parameter logistic item response model using the R package TAM (Kiefer, Robitzsch, & Wu, 2014; R Core Team, 2014). Only the knowledge acquisition score was used for analyses as it is conceptually closer to the assumed role of problem solving in an ICT context.

As to our knowledge no test exists that covers the comprehension of textual and graphical elements, we decided on a well-proven time limited reading comprehension test. Test-takers needed to complete gaps in a text by choosing the most fitting out of three presented words, and were expected to do this within a limit of four minutes. According to the authors, the comprehension score had a retest-reliability of $r = .87$ (LGVT; Schneider, Schlagmüller, & Ennemoser, 2007). Comprehension scores were built based on the sum of correct solutions, including a penalty for incorrectly solved items and were standardized for analyses.

ICT specific knowledge was assessed using a subscale that assesses theoretical computer knowledge from the Computer Literacy Inventory (INCOBI-R; Richter, Naumann, & Horz, 2010). This scale consists of 20 multiple choice questions asking for different terms concerning computers. The correct answer had to be selected out of four alternatives. The sum of correctly answered questions was counted and standardized for analyses. For this sample, Cronbach's alpha was $\alpha = .68$.

Task Characteristics

The indicator for reading load was the number of words on relevant pages for task solution and was counted for all 64 items, including the instructions ($M = 235.9$, $SD = 250.3$, $Min = 45$, $Max = 1815$). Instructions consisted on average of $M = 53.4$ words ($SD = 12.4$). The indicator for the intrinsic complexity of the task was the minimum number of required user interactions to solve the item correctly. The amount of required user interactions was counted for each item based on the expected solution. Only diverse but no iterative behavior was counted ($M = 6.0$, $SD = 3.4$, $Min = 1$, $Max = 16$). As iterative behavior is not assumed to increase the requirement of problem solving skills, opening (for instance) five emails in a row was only counted as one required interaction. Both variables were standardized for analyses.

Data Analyses

We applied generalized linear mixed models (GLMM; De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx, & Partchev, 2011; Wilson, De Boeck, & Carstensen, 2008) and the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014). This way, relations can be investigated as fixed and random effects. Fixed effects assume a constant relation across all items, and random effects allow variation, assuming a different relation for different items. The probability of solving an item correctly is expressed by the logit of the probability for one person (p) solving an item (i) correctly (P_{pi}). As the probability of task success is explained, higher values can be interpreted in terms of item easiness.

$$ln\left[\frac{P_{pi}}{1-P_{pi}}\right] = \beta_0 + \sum_{v=1}^{V} \beta_{1v} X_{(p,i)v} + b_{0i} + b_{0p} + b_{0s} \quad (1)$$

Equation 1 describes the model used to analyze Hypothesis 1, which was extended for the other analyses. An overall intercept (β_0), random effects across items (b_{0i}), persons (b_{0p}) and schools (b_{0s}) were modeled and all three person variables v (problem solving, reading comprehension, computer knowledge) were included as fixed effects (β_{1v}). To allow varying relations of the person variables across items (Hypothesis 2), a full random effects model was applied. The effects of the three variables were allowed to vary across items, as well as to correlate with item easiness and among each other. Whether these variations contributed significantly to model fit was investigated by comparing this model to the model of Hypothesis 1. For Hypothesis 3, interaction terms of variables and task characteristics were added, to investigate whether the relation of the component skills to ICT skills was stronger for items with more indicators.

Results

According to *Hypothesis 1*, all three person variables had unique positive effects on task success across all ICT items (Table 1; problem solving: $\beta = 0.32$, $p < .001$, reading comprehension: $\beta = 0.17$, $p < .001$, computer knowledge: $\beta = 0.18$, $p < .001$). Including these variables explained 35.4 % of person variance and 71.1% of school variance.

--- Insert Table 1 about here ---

The full random effects model (Table 2) fitted the data better than the model without variation (χ^2 (9) = 19.98, $p = .018$), indicating varying effects across items around the fixed effects. The variation of the effects of all three variables across items is visualized in Figure 2, sorted by item difficulty. The random effects are displayed as variation from their fixed effect. Problem solving was positively related to task success for all items, while the relation of computer knowledge to task success was around zero for a few easy items, as was the relation of reading comprehension for a few medium and very difficult items (*Hypothesis 2*). The effect of computer knowledge tended to be higher for more difficult items ($r = -.63$). The effect of generic skills tended to be higher for easier items (RC: $r = .50$; PS: $r = .15$).

--- Insert Table 2 and Figure 2 about here ---

Item characteristics, which were assumed to evoke reading and problem solving processes, were included in the analyses (*Hypotheses 3*; Table 3). As expected, the intrinsic complexity of a task interacted positively with the effect of problem solving ($\beta = 0.11$, $p = .016$). Including this item-level predictor accounted for 37.3% of the variance of the random problem solving effect across items (furthermore, 10.4% of the variance in item easiness was explained). Contrary to expected outcomes, reading load did not interact with the effect of reading comprehension ($\beta = 0.02$, $p = .670$).

--- Insert Table 3 about here ---

Discussion

Across all items, all three variables - reading comprehension skills, problem solving skills and computer knowledge - predicted task success positively (Hypothesis 1). Problem solving, as a generic skill, was the strongest predictor for task success across all items (fixed effects in Figure 2) and was also (random effects in Figure 2) for most items more predictive on item level than the ICT-specific skill, computer knowledge. We suggest that this might be caused by the operationalization of the tests. The problem solving test requires interacting dynamically with a surface, which might have caused the higher correlations for problem solving and the performance-based ICT skills items. The computer knowledge test, in contrast, required only few interactions with the environment and can be even administered in a paper-based fashion. Taken together, the three variables explained a substantial amount of person variance (35.4%) in solving ICT skills items.

Construct interpretation (as defined in Figure 1) was supported on the whole (Hypothesis 2), as computer knowledge and at least one of the generic skills were required for most items. However, the item-specific effects of computer knowledge were around zero for a few easy items. For those items, construct interpretation might be not valid, because the required knowledge did not differentiate between persons. Those items should be double-checked or excluded. For a few harder items, the effect of reading comprehension was around zero. As problem solving was still required for solving these items, construct interpretation is still valid.

The relations of generic skills and ICT skills were assumed to be caused by similar cognitive processes that were triggered by task characteristics. As expected, the required user interactions for item solution explained the relation of problem solving, while the amount of words per item could not explain the relation of reading comprehension (Hypothesis 3). Two

explanations are possible. At first, the validity of the test score interpretation could be questioned, because it is possible that the correlation of reading comprehension and ICT skills was not caused by similar processes performed in ICT skills and reading tasks. If the relation of reading was caused instead by third variables rather by characteristics of the task, the definition according to Figure 1 would not be true for the ICT skills items, as reading processes would not be performed in the items. However, the varying relations of reading comprehension across items spoke against this interpretation, as they indicated that the relation of reading depended on properties of the items and makes the following explanation more plausible. A second explanation would question the number of words as a reliable indicator for reading load, as they were also not at all related to item difficulty (cf. Table 3). Is it possible that not every word on a page has to necessarily be read to solve the items correctly. The next steps should be to analyze in separate studies which words or elements are processed in items, in order to detect mandatory elements for task solution (for instance, via eye tracking methods) and to check whether a better indicator can be established.

Considering all these results together, the validity of the construct interpretation of the test score is generally supported for everyday ICT tasks performed by 15 year old students in Germany. Still, a few easy items should be re-checked or excluded, as ICT specific skills were not decisive for this sample to solve these items.

Conclusion

Within this study we extended the classical nomothetic span approach and investigated relations to other variables on item level as well as interaction effects with task characteristics. Significant varying relations across items supported the appropriateness of this approach. Beyond that, we demonstrated a bridge between a domain that is rather operationally defined, namely ICT skills, to well-established and well-studied constructs. Henceforth, we overcame

shortcomings of recent conceptualizations by providing detailed suggestions concerning the nature and reasons for relations, allowing for the collection of validity evidence for the construct interpretation of test scores from the ICT skills test based on relations to other variables.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, URL http://CRAN.R-project.org/package=lme4.

Binkley, M., Erstad, O., Herman, J., Raizen, R., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills (pp. 17 – 66)*. Dordrecht: Springer.

Calvani, A., Cartelli, A., Fini, A., & Ranieri, M. (2009). Models and instruments for assessing digital competence at school. *Journal of e-Learning and Knowledge Society-English Version*, *4*, 183–193. Retrieved from http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/288/270.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197. http://dx.doi.org/10.1037/0033-2909.93.1.179

Engelhardt, L., Goldhammer, F., Naumann, J. & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. *Computers in Human Behavior, 76*, 683-692.

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia*, *13*, 93-107. Retrieved from http://www.openu.ac.il/Personal_sites/download/Digital-literacy2004-JEMH.pdf

European Parliament and the Council (2006). Recommendation of the European Parliament and the Council of 18 December 2006 on key competences for lifelong learning. *Official Journal of the European Union, L394*. Retrieved from http://www.alfa-trall.eu/wp-content/uploads/2012/01/EU2007-keyCompetencesL3-brochure.pdf

Ferrari, A., Punie, Y., & Redecker, C. (2012). Understanding digital competence in the 21st century: an analysis of current frameworks. In *21st Century Learning for 21st Century Skills* (pp. 79-92). Springer Berlin Heidelberg. doi:10.1007/978-3-642-33263-0_7

Fraillon, J., & Ainley, J. (2010). The IEA international study of computer and information literacy (ICILS). Retrieved from http://www.researchgate.net/profile/John_Ainley/publication/268297993_The_IEA_International_Study_of_Computer_and_Information_Literacy_%28ICILS%29/links/54eba4330cf2082851be49a9.pdf.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment*, *29*, 263-275.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of Complex Problem Solving: concept, implementation, and application. *Educational Technology Research and Development, 61*, 407-421.

International ICT Literacy Panel (2002). Digital Transformation: A Framework for ICT Literacy. *Educational Testing Service*. Princeton, NJ. Retrieved from http://www.ets.org/Media/Research/pdf/ICTREPORT.pdf.

Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: An R Package for Item Response Modelling.*

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. *Handbook of educational psychology, 2*, 287-303.

OECD (2012). *Literacy, Numeracy, and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills, OECD Publishing.*

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Richter, T., Naumann, J., & Horz, H. (2010). Das Inventar zur Computerbildung (revidierte Fassung*)* [A Revised Version of the Computer Literacy Inventory]. *Zeitschrift für Pädagogische Psychologie*, *24*, 23-37.

Rölke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012* (pp. 344-353). Chesapeake, VA: AACE.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12: Lesegeschwindigkeits- und –verständnistest für die Klassen 6-12* [Reading Speed and Comprehension Test for Grades 6 to 12]. Göttingen: Hogrefe.

Schnotz, W. (2005). An Integrated Model of Text and Picture Comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–69). New York, NY, US: Cambridge University Press.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review, 19*, 58-84.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist, 26*, 145–159. http://dx.doi.org/10.1037/h0030806

Wenzel, S.F.C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F., Naumann, J., & Horz, H. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) (CavE-ICT) [Computer-based, adaptive and behavior-related assessment of information and communication-related competencies (ICT skills)]. In BMBF (Hrsg*.). Forschung in Ankopplung an Large-Scale Assessments* (pp. 161-180). Bonn, Berlin: BMBF.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91-120). Göttingen: Hogrefe.

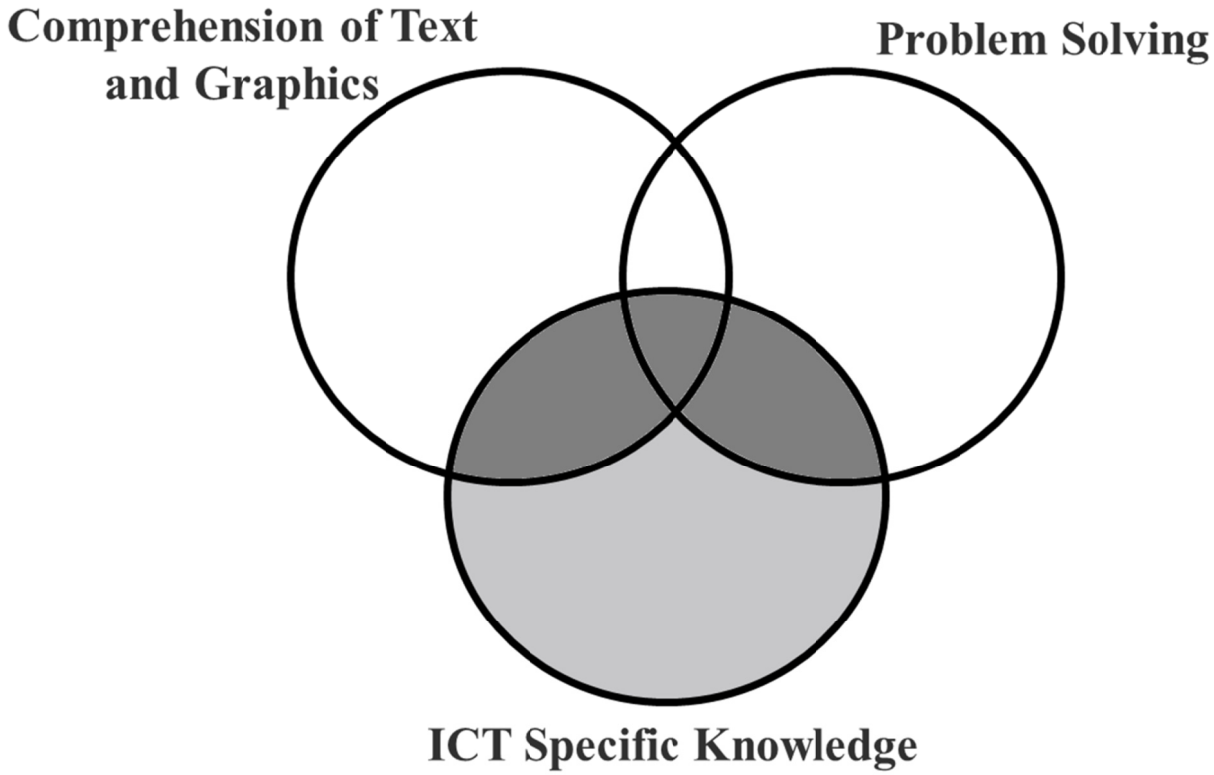


Figure 1. Definition of ICT skills. Adapted from Wenzel et al. (2016, p. 164).

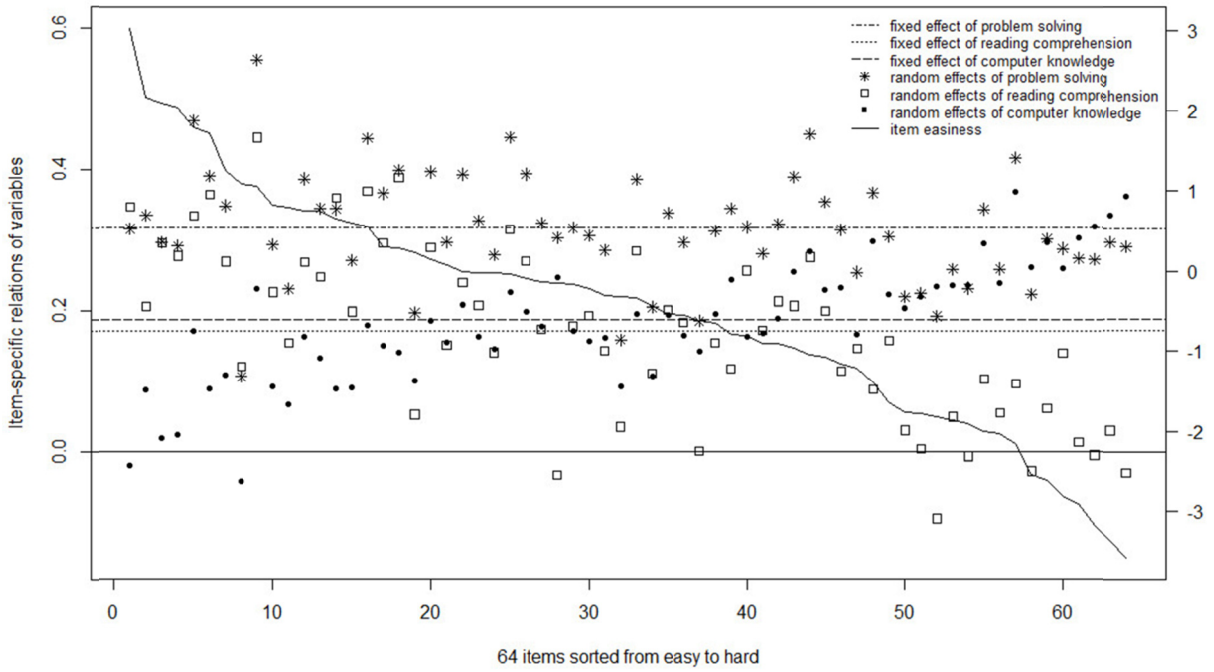


Figure 2. Visualized effects based on Table 2. Y-axis on the right side relates to item easiness.

Table 1

Hypothesis 1. Relations of person variables across all items.

	Baseline Model					Hypothesis 1				
Parameters	σ^2	β	*SE*	*z*	*p*	σ^2	β	*SE*	*z*	*p*
Fixed effects										
Intercept		-0.43	.22	-1.98	.048		-0.51	.20	-2.49	.013
Problem Solving							0.32	.05	6.56	<.001
Reading Comprehension							0.17	.05	3.51	<.001
Computer Knowledge							0.18	.05	3.48	<.001
Random effects										
School	0.25					0.07				
Person	0.39					0.25				
Item	2.34					2.33				

Note. Baseline Model: value ~ (1| item) + (1 | person) + (1 | school); Model for Hypothesis 1: value ~ PS+RC+CK+ (1| item) + (1 | person) + (1 | school); persons =256, schools = 34, items= 64.

Table 2

Hypothesis 2. Full random effects model with person variables varying across items, co-variating with item easiness and among each other.

Parameters	σ^2	r_i	r_{ps}	r_{rc}	β	*SE*	*z*	*p*
Fixed effects								
Intercept					-0.50	.21	-2.46	.014
Problem Solving					0.32	.05	6.01	<.001
Reading Comprehension					0.17	.06	3.10	.002
Computer Knowledge					0.19	.05	3.44	<.001
Random effects								
School	0.07							
Person	0.25							
Item	2.34							
Item – Problem Solving	0.02	.15						
Item – Reading Comprehension	0.03	.50	.58					
Item – Computer Knowledge	0.01	-.63	.54	-.29				

Note. Model: value ~ PS + RC +CK+ (1 + PS + RC + CK | item) + (1 | person) + (1 | school); persons =256, schools = 34, items= 64; Optimizer: "bobyqa". r_i= random effect correlation with item easiness; r_{ps}= random effect correlation with problem solving; r_{rc} = random effect correlation with reading comprehension.

Table 3

Hypothesis 3. Full random effects model and interaction effects between person variables and task characteristics.

Parameters	σ^2	r_i	r_{ps}	r_{rc}	β	*SE*	*z*	*p*
Fixed effects								
Intercept					-0.47	.20	-2.41	.016
Problem Solving (PS)					0.33	.05	6.36	<.001
Reading Comprehension (RC)					0.17	.06	3.09	.002
Computer Knowledge (CK)					0.18	.05	3.354	<.001
Intrinsic Complexity					-0.41	.19	-2.11	.034
Intrinsic Complexity: PS					0.11	.04	2.41	.016
Reading Load					-0.17	.15	-1.15	.249
Reading Load: RC					0.02	.04	0.43	.670
Random effects								
School	0.07							
Person	0.25							
Item	2.10							
Item – Problem Solving	0.01	.26						
Item – Reading Comprehension	0.03	.53	.77					
Item – Computer Knowledge	0.01	-.28	-.33	-.35				

Note. Model: value ~ PS* Intrinsic Complexity + RC*Reading Load + CK + (1 + PS + RC + CK | item) + (1 | person) + (1 | school); persons = 256, schools = 34, items = 64; Optimizer: "bobyqa". r_i= random effect correlation with item easiness; r_{ps}= random effect correlation with problem solving; r_{rc} = random effect correlation with reading comprehension.

Anhang C

Arbeit 3: Engelhardt, L., Goldhammer, F., Naumann, J., & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. *Computers in Human Behavior, 76,* 683-692. doi: 10.1016/j.chb.2017.02.020

Computers in Human Behavior 76 (2017) 683–692

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Experimental validation strategies for heterogeneous computer-based assessment items

Lena Engelhardt [a, *], Frank Goldhammer [a, b], Johannes Naumann [c], Andreas Frey [d, e]

[a] German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany
[b] Centre for International Student Assessment (ZIB), Germany
[c] Goethe University Frankfurt am Main, Germany
[d] Friedrich Schiller University Jena, Germany
[e] Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

ARTICLE INFO

Article history:
Received 17 November 2016
Received in revised form
30 December 2016
Accepted 5 February 2017
Available online 5 February 2017
A previous version of this article was presented on the National Council on Measurement in Education (NCME) 2016.

Keywords:
Validation
Experimental strategies
Heterogeneous item sets
Computer-based assessment
ICT skills

ABSTRACT

Computer-based assessments open up new possibilities to measure constructs in authentic settings. They are especially promising to measure 21st century skills, as for instance information and communication technologies (ICT) skills. Items tapping such constructs may be diverse regarding design principles and content and thus form a heterogeneous item set. Existing validation approaches, as the construct representation approach by Embretson (1983), however, require homogenous item sets in the sense that a particular task characteristic can be applied to all items. To apply this validation rational also for heterogeneous item sets, two experimental approaches are proposed based on the idea to create variants of items by systematically manipulating task characteristics. The *change*-approach investigates whether the manipulation affects construct-related demands and the *eliminate*-approach whether the test score represents the targeted skill dimension. Both approaches were applied within an empirical study ($N = 983$) using heterogeneous items from an ICT skills test. The results show how changes of ICT-specific task characteristics influenced item difficulty without changing the represented construct. Additionally, eliminating the intended skill dimension led to easier items and changed the construct partly. Overall, the suggested experimental approaches provide a useful validation tool for 21st century skills assessed by heterogeneous items.

Assessments are increasingly carried out by means of computers enabling the automatic evaluation of responses, and more efficient (i.e., adaptive) testing. With the advance of computer-based assessment, there is an ongoing and pertinent debate around the validity of the test score interpretation, as computer skills are required to complete the tasks. This is true even in domains where it would appear naturally at first sight to use the computer as an assessment tool, because the targeted skill unfolds in a digital environment as well, such as digital reading (see OECD, 2011). Even in these domains, extra care needs to be taken that the assessment targets individual differences in reading-related processes, and not merely computer skills. Actually, for most so-called 21st century skills (e.g., problem solving, collaboration, information literacy; Binkley et al., 2012) computers are needed in order to measure them in realistic settings and specifically, simulated environments provide more authentic task settings.

Also educational large-scale-assessments such as PISA (Programme for International Student Assessment; OECD, 2014) or PIAAC (Programme for the International Assessment of Adult Competencies; OECD, 2012) are nowadays assessed by means of computers. Computer-based assessments allow assessing skills performance-based, by not only asking for instance “how good are you in digital reading?” but asking students to actually perform digital reading tasks. Besides, test scores from such studies are interpreted more general in terms of requirements for societal participation (e.g., in PISA). They are not in the first place based on conventional psychological constructs such as intelligence, but on ”institutionally defined knowledge domains” (Watermann & Klieme, 2002, p. 2). To be able to justify such a far ranging test score interpretation, typically broad constructs and – in turn – heterogeneous items representing a wide range of contents and

* Corresponding author. German Institute for International Educational Research (DIPF), Schlossstraße 29, 60486 Frankfurt am Main, Germany.
E-mail address: lengelhardt@dipf.de (L. Engelhardt).

http://dx.doi.org/10.1016/j.chb.2017.02.020
0747-5632/

situations are needed. Items in these educational studies differ more strongly from each other than items used to assess conventional psychological constructs. This is because items are often instructed and designed in a contextualized way within a certain situation. They differ in their appearance, but also in the demands and in the knowledge they require (for sample questions see e.g., https://www.oecd.org/pisa/test/form/). With „heterogeneous", we refer to a property of items measuring a certain construct, but not to assumptions regarding the underlying dimensional structure. An example for such a broadly defined but one-dimensional 21st century skills construct is computer and information literacy assessed in the computer-based ICILS-Study (International Computer and Information Literacy Study; Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

Such "innovative item formats" have obvious advantages in terms of construct representation (Sireci & Zenisky, 2006, p. 329) because items can be then more contextualized or authentic, but give rise to new challenges for the validation of test score interpretations (e.g., Linn, Baker, & Dunbar, 1991), as further skills, for instance skills to interact with a computer environment, are involved in the task solution process especially in performance-based assessments. Thus, validation needs to take this into account by providing evidence that the assumed construct-related processes are actually exercised by the test-taker. The validity-threatening potential of such skills is even more an issue when domains are being assessed with the computer that by themselves have no overlaps with ICTs, such as print reading, science, or mathematics. In traditional correlational approaches, these validity threats are addressed by including additional measures in the validation design that directly assess computer or ICT skills. Thereby, discriminant evidence can be provided (see AERA, APA, & NCME, 2014).

The goal of this paper is to present two experimental validation strategies that offer an additional way of dealing with the issue of validity in computer-based assessment and can be also applied to heterogeneous item sets. In the following section, we first briefly describe Embretson's (1983) construct representation approach as the two suggested approaches are based on this validation rationale. We refer then to how the two experimental validation strategies proposed in this article differ from Embretson's approach. They are described in term of their conceptual basis and also in terms of concrete consequences for building hypotheses in the validation process. An application of the two approaches is presented using empirical data gathered with a test measuring information and communication technology (ICT) skills.

1. Embretson's construct representation approach

Validity is not a property of a test but of "the interpretations of test scores for proposed uses" (cf. AERA et al, 2014, p.11). Kane (2013) suggests that especially a theory-based interpretation, for instance related to the construct, requires ambitious claims of validity. Thus, the strategy for validating a test score interpretation depends on the intended use and the inferences that should be made based on the test scores. A very important and also ambitious claim for justifying the construct interpretation would refer to the relation of task characteristics to the test-taker's score based on the underlying process model that is derived from theory (Kane, 2013). Such claims can be investigated using Embretson's construct representation approach.

The rationale behind the construct representation approach is to determine task characteristics that should theoretically evoke the targeted cognitive processes. These task characteristics – that should also have guided the item development process – are then related to task performance, for instance to item difficulty. If items showing those task characteristics to a greater extent are also harder, test scores can be interpreted as determined by the targeted construct. An example for such a task characteristic could be the number of transformations in a mental rotation task that describes the items' complexity (cf. Embretson, 1983) or the number of orthographic neighbors in a word recognition task, thus words that differ in their spelling from the targeted word in only one letter. Such task characteristics can be described as complexity factors that can be quantified and describe the complexity of an item in terms of cognitive processes that have to be performed. Thus, the approach refers to the cognitive processes that are assumed to occur while working on the task.

The construct representation approach was applied in many studies, for instance for mental rotation tasks (Caissie, Vigneau, & Bors, 2009), problem solving tasks (Greiff, Krkovic, & Nagy, 2014; Stadler, Niepel, & Greiff, 2016), or reading comprehension tasks (Hartig & Frey, 2012). Note that in these studies, the stimulus material was homogeneous, that is all items could be described by the same stimulus characteristics in the items. In the mental rotation task, for instance, stimuli have to be evaluated whether they represent a rotation of the initial figure or not, which leads to items with comparable stimulus materials and task solution processes. Defining comparable task characteristics across items, however, might be only feasible in more restricted domains that are not as broad as some domains that are assessed in large-scale assessments (Watermann & Klieme, 2002). This holds for instance for the ICT skills test used in this study, because users have to deal with, for instance, different applications (e.g., browser or e-mail) and different information tasks (e.g., access or evaluate information).

Combining one type of information tasks with one environment might in fact compose a facet of ICT skills that can be measured with a homogeneous items set, making it possible to employ Embretson's (1983) construct representation approach. For example, Pfaff and Goldhammer (2011; see also Hahnel, Goldhammer, Naumann, & Kröhne, 2016) described a test measuring the evaluation of information presented in browser environments. In this case, item features can be identified that are comparable across all items, as for instance the number of to-be-accessed hypertext pages. However, when a comprehensive assessment of ICT skills is intended, the different information tasks and the different applications imply that it will be difficult to find task features that can be defined for all tasks in the assessment alike. Think, for instance, of a task requiring information to be created using computers. Such a task might require changing font sizes or the position of text fields in a presentation. A task requiring information to be accessed in contrast might require the test-taker rather to navigate text presented in a browser environment. The two suggested experimental approaches make the rational of relating task characteristics to item difficulty also feasible for heterogeneous item sets.

Embretson (1983) describes besides "construct representation" a second approach to validation, the "nomothetic span" approach. While "construct representation" focuses on task differences, "nomothetic span" targets individual differences. In the nomothetic span approach, the relations to other constructs as predicted by the nomological network or that are supposed to underlie the item solution process are investigated. The idea of a nomological network is to find evidence for supporting the targeted test score interpretation for a specific use by investigating the relation to other variables. These can be variables that are assumed to be related (convergent evidence) and variables that are assumed not to be related to the test scores (discriminant evidence) (cf. AERA et al, 2014).

2. Two new experimental validation strategies

We want to apply Embretson's approach (1983) of relating task characteristics to task performance also to heterogeneous item sets, where potentially every item belongs to a separate item type. Such item sets are frequently used in computer-based (large-scale) assessments of student achievement to measure broadly defined constructs. The novelty of the two proposed experimental approaches is to systematically construct variants of original items by manipulating certain task characteristics for validation purposes. In a homogeneous item set, these variants already exist, as all items are of the same type. Two mental rotation tasks, for instance, will differ in the number of rotations they require, and nothing more. In heterogeneous item sets, in contrast, two items will differ in the features that characterize the task, making it difficult to pinpoint which item characteristic might drive differences in item difficulty. The general idea behind the two new experimental validation strategies is thus to deliberately manipulate individual characteristics of existing items. These manipulations are such that from the construct definition it can be expected that either the manipulated item is easier or harder than the original one (*change*-approach), or taps a different construct (*eliminate*-approach).

Four different analyses are required to investigate whether the two manipulations affected task performance as expected (Table 1) and will be described more detailed using the example of an ICT skills item (Fig. 1).

To solve this item, the test-taker has to decide for each e-mail in his e-mail inbox whether it is relevant for a new colleague. If the user decides for relevance, he needs to forward the e-mail to the address that is provided in the instruction. The crucial aspect in this task is whether the third e-mail is identified correctly as a hoax e-mail that should not be forwarded.

2.1. Change-approach

The *change*-approach is based on the construct representation approach, in which item characteristics are related to item difficulty. But here, these characteristics are not identified for all items, but changed by developing a variant for a particular item where exactly this characteristic is changed. *Change* refers to a change of item-specific task characteristics that are assumed to evoke the construct that is supposed to cause differences in the test score. In terms of the information-processing paradigm, *change* refers to a change of the cognitive process. The task solution should be easier or harder depending on the direction in which the processes are changed. A *change*-variant of the example item (Fig. 1) can be created through changing the easiness to detect the third e-mail as a hoax e-mail. This aspect is crucial to the item as it requires ICT-specific evaluation skills. Since the presumed author of this e-mail is a rather trustworthy source, namely a colleague, the trustworthiness can be decreased in the *change*-variant by introducing an unknown author (a mailing list), potentially to the effect that the e-mail is read and evaluated more critically. If indeed the authorship serves as a criterion for evaluating e-mails, this item variant should be easier.

These considerations have two implications (cf. Table 1) for the functioning of changed items. First, depending on the nature of the change, the changed item should be easier or harder than the original. Second, the relations to other constructs should not be affected, as despite being easier or harder to perform, the cognitive processes required by an item (e.g., evaluating the e-mails) stay the same.

Previous studies already varied task characteristics in homogeneous item sets, for instance in matrices tasks. They followed predefined construction rules across all items and the purpose was for instance item writing (Hornke & Habon, 1986). In a matrices task, all items belong to the same item type because each item asks the test taker to identify a missing peace by applying different rules (e.g. addition). Although, the type of rules to be applied may differ across items, still each item is characterized by the requirement to apply one or more rules. We thus describe such item sets as homogeneous. We see the difference and innovation of the *change*-approach in that it can be also applied to heterogeneous item sets and that the purpose is in first line for validation but not for constructing new items. Other studies which were concerned with validation and also had to deal with heterogeneous item sets, related instead of task characteristics expert ratings, for instance regarding the cognitive demands, to item difficulty (e.g. Watermann & Klieme, 2002). With the *change*-approach we suggest to manipulate those task characteristics (e.g. trustworthiness in the given example), which are assumed to require these cognitive demands. Since these manipulations can be made for every item separately, it does not matter how close and homogeneous items are to each other and whether comparable task characteristics can be found across items. Furthermore, while rating the cognitive demands delegates the validation process to the experts, manipulating task characteristics involves the test-taker stronger into the validation process. Similar as in earlier approaches of item manipulations (e.g., Hornke & Habon, 1986), also more than one *change*-manipulation could be possibly applied to one item, since items can be made easier or harder and also the degree of manipulation can vary.

2.2. Eliminate-approach

The *eliminate*-approach is based on investigating the nomothetic span. The relation to other variables being part of the assumed nomological network is evaluated for eliminate and original items. It is important to note that this approach is not primarily meant to investigate whether the relation exists as predicted by a nomological network, but goes further and compares the relations for manipulated and original items in order to investigate whether a change in the task characteristics affects the relation to other variables as expected. *Eliminate* refers to the elimination of all task characteristics that represent the construct, that is supposed to cause individual differences in the test score. Described in terms of the information-processing paradigm, elimination refers to the entire removal of the need to perform a specific cognitive process. *Eliminate*-items were created through elimination of the requirement to apply higher order ICT-skills involving judgement and decision. Thus, *eliminate*-items only required test-takers to perform basic operations, such as clicking buttons.

Table 1
Analyses for the experimental approaches.

	Indicators for task performance	
Manipulation	Item difficulty	Relation to construct-related variables
Changing (specific aspects)	H1a: Easier or harder (than original items)	H1b: Same pattern (as original items)
Eliminating (a whole skill dimension)	H2a: Easier (than original items)	H2b: Different pattern (than original items)

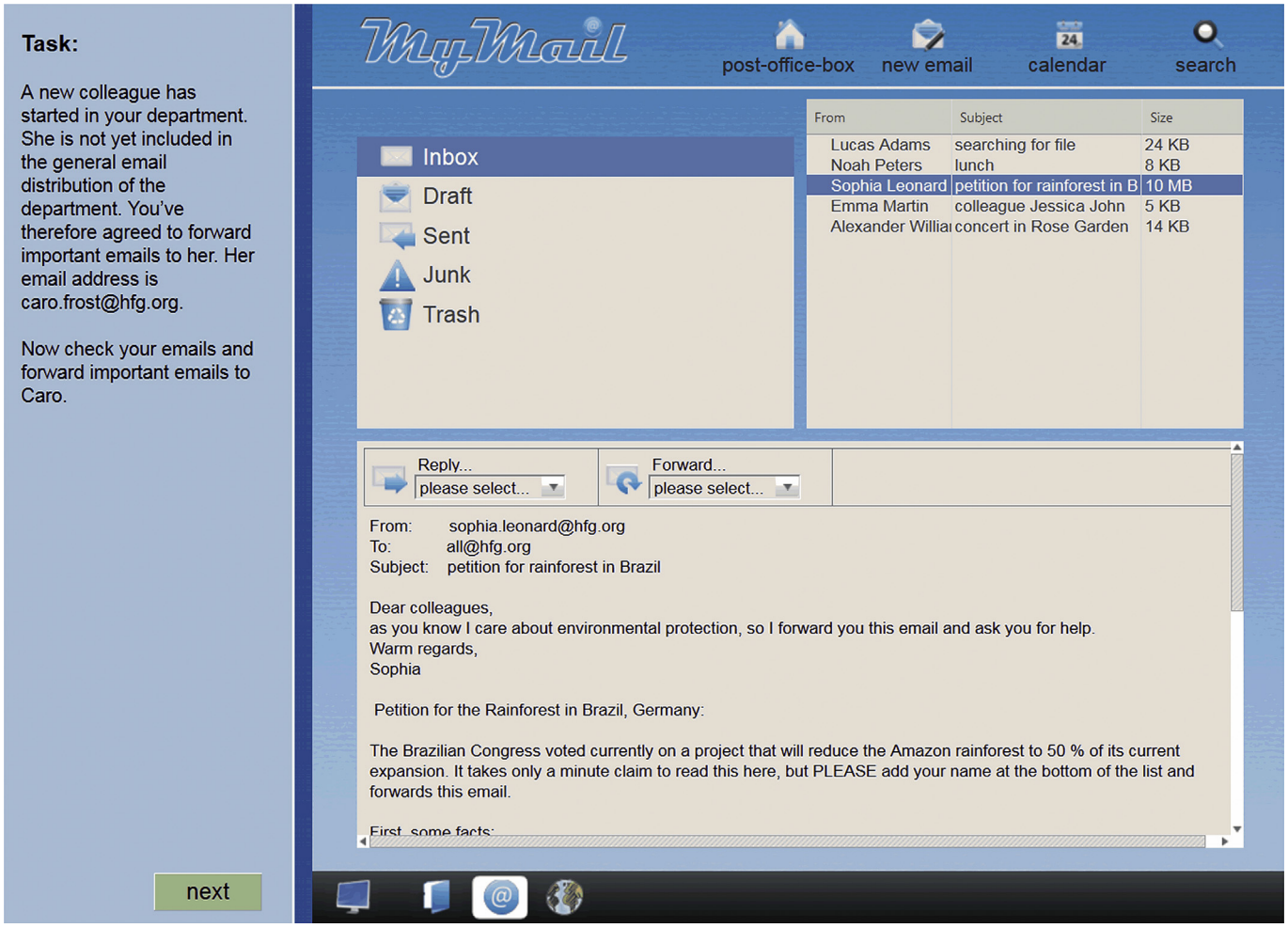


Fig. 1. Example of an ICT skills test item.

Through this, presumably the nature of the targeted construct was changed. In the example item, the correct e-mail that needed to be forwarded was already mentioned in the instruction. The last sentence of the instruction (Fig. 1) "Now check your e-mails and forward important e-mails to Caro" was modified into "Now check your e-mails and forward the e-mail of Emma Martin to Caro". Again, these considerations have implications for the likely functioning of *eliminate*-items, as compared to the original item they were derived from. First, the probability of solving the item should be increased, if the requirement of performing a specific cognitive operation is removed from the item. Second, other than *change*-items, the correlations of eliminate-items to other variables should be affected, as removing the requirement to perform a specific cognitive process from an item will by definition change the nature of the construct assessed by the item.

We see the advantage of the eliminate-approach in constructing item-variants that lack the targeted skill dimension to investigate whether, besides item difficulty, the measured construct changes. This might seem to be not reasonable on the first sight, since these items can obviously not be used in further assessments. However, generally speaking correlation does not imply causation. Thus, for instance even if a computer-based reading test showed a strong correlation with a paper-based reading test, but not with a test of ICT skills, there would be always interpretations other than the intended (e.g., there is a common underlying ability, that "causes" the performance in both the computer and paper-pencil test of reading). In contrast, when *eliminate*-items are being administered to subjects randomly, the changes in item difficulty can be causally attributed to the manipulations in the items. Thus, the *eliminate*-approach allows to challenge seriously (Kane, 2013, p. 15) the assumption that correlations of test scores with related variables are caused by the assumed skill dimension. It might seem to be trivial that a correlation changes once the targeted skill dimension is eliminated, but it is not trivial in items that are very complex and require for instance also some navigation or reading skills to read the instruction. For example, if the relation of test scores from the ICT skills items to ICT related variables changes by eliminating the evaluation process from the item, it is supported that the relation was indeed caused by the required evaluation process.

2.3. Comparing the two approaches

How do the two strategies, *eliminate* and *change*, relate to each other? On a conceptual level, both manipulations differ in how they affect the cognitive processes while solving an item. The *change*-approach only affects the targeted cognitive process gradually (e.g., how difficult the evaluation process is), by making this process easier or harder. The *eliminate*-approach, however, would eliminate all targeted cognitive processes (e.g., no evaluation process is required) that belong to the targeted skill dimension. This is why even if the evaluation of the hoax e-mail became rather easy, there will be still some evaluation skills needed in a *change*-item to treat

this email correctly, but not in an *eliminate*-item. Thus, *eliminate*-manipulations can only lead to easier items, because cognitive processes are removed from the solution process, while *change*-manipulations can change difficulties in both directions and in different intensity. As a consequence, only one *eliminate*-manipulation can be carried out per item, while several manipulations are possible for *change*-items.

At second, the manipulations are carried out addressing different part of the items: *Change*-manipulations are carried out by changing task characteristics within the item, for instance the author of an e-mail, while *eliminate*-manipulations are carried out by adding information to the instruction. This is why an *eliminate*- and a *change*-item will never be the same although they may both decrease item difficulty.

And finally, they differ regarding the effect they are intended to have on construct-related variables. The *change*-manipulation leads to items that are intended to measure still the same construct and differ only in their difficulties, while the *eliminate*-manipulation leads to items that should not measure anymore the same construct. As a consequence, *change*-items can be also used for eventual testing since they should measure the same construct, while *eliminate*-items cannot. Such *change*-variants can be useful for assessing specific samples, for instance regarding age or skill level, or for adaptive tests, where items with difficulties across the whole ability range are needed.

2.4. Theoretical and practical gains of the experimental approaches

One advantage of these procedures compared to correlational approaches (e.g. investigating the nomological network) is that potentially (several) confounding variables must not all be added to the validation design (although this of course comes at the price that the manipulated items need to be included in the assessment). With these procedures, only one construct-related variable can be used to investigate at first whether the expected relation actually exists (convergent evidence), and also at second and third, whether the relation to *change*- and *eliminate*-items changes or not.

When *change*- or *eliminate*-items are being administered to subjects randomly, the changes in test scores can be causally attributed to the changes in the items. By these means, the *change*- and *eliminate*-approach also add to the validity argument by addressing the cognitive processes that are assumed in a given item (AERA et al, 2014).

We consider especially the combination of both approaches as promising. The results of the *eliminate*- and *change*-approach strengthen each other: A relation that changes by manipulating task characteristics (*eliminate*-approach) supports that it is not trivial that the relation to *change*-items is the same after manipulating task characteristics, and vice versa. As positive side-effect, both approaches require considering the validation strategies already in the process of item development, which can be beneficial if the changes are already planned together with the item construction also for the original versions of the items. Additionally, producing item variants takes only low effort once the original item is developed. This is not negligible, since implementing authentic items on a computer can be rather effortful and make feasible validation strategies even more important.

3. Applying the experimental approaches to the construct 'ICT skills'

3.1. Construct representation of ICT skills

In this research, we apply the *change*- and *eliminate*-approaches to validation to a test of ICT skills: ICT skills form a prototypical instance of a competence that is so broadly defined it can hardly be measured using a homogeneous set of items.

Different conceptualizations of ICT skills focus on different skill levels, such as basic computer skills (Goldhammer, Naumann, & Keβel, 2013), cognitive skills when using ICT (Eshet-Alkalai & Chajut, 2010), or the interplay of different levels of skills in one task (van Deursen & van Dijk, 2009). We focus on higher-order skills. Thus, we do not target basic ICT tasks that can be routinely performed on the basis of a pre-defined sequence of clicks. Rather, we target skills in such a way that they involve components of judgement and decision making (see the example item). A test measuring basic ICT skills might present test-takers with an e-mail, and then requiring them to enter a given address in the address field of some e-mail client, find, and click the "forward"-button. In contrast, higher-order ICT skills as addressed here would include a decision about whether a given e-mail should be forwarded to a given person or number of persons in a given situation. These decisions should be based on previous experiences, because experiences with ICT seem to determine skills (Eshet-Alkalai & Chajut, 2010). The decisions should be also based on knowledge specific to the ICT domain (henceforth "technical knowledge"), which is part of several ICT conceptualizations (Fraillon & Ainley, 2010; International ICT Literacy Panel, 2002; van Deursen & van Dijk, 2009). In our example, identifying the third e-mail correctly as a hoax (cf. Fig. 1) requires not only reading skills to understand the purpose of the e-mail but also evaluation skills in order to decide not to follow the call in the e-mail to forward. This e-mail is sent by a colleague who might be regarded as a trustworthy source (cognitive authority; Rieh, 2002). For this decision, higher-order ICT specific skills are needed that are based on knowledge and experience about typical markers of spam.

3.2. Developing a heterogeneous item set

In ICT environments, tasks can pose widely different cognitive challenges, or require different cognitive operations. For instance, a task might require a person to either access, manage, integrate, evaluate, or create information (International ICT Literacy Panel, 2002). In addition, the environment in which a task occurs can differ widely across tasks, and employ tools such as spreadsheets, browsers, e-mail clients, text-processors, etc. Thus, through the combination of ICT task and various environments, items are even within one cognitive operation heterogeneous. For instance, evaluate tasks may not only require to consider information regarding the author but other criteria of truth as well (Rieh, 2002). But also regarding the relevance of websites (Pfaff & Goldhammer, 2011), or the estimated value of information (Whittaker & Snider, 1996). If these different aspects of evaluating information, besides the other information tasks are included into the test, comparable and quantifiable criteria cannot even be found within all evaluate items. Although the construct is measured by heterogeneous items, we still assume that the construct of ICT skills is needed to solve all these items (i.e., assumption of one-dimensionality).

3.3. Hypotheses

Following the general steps for the *change*-approach, we expected the following (cf. Table 1): Changing task characteristics has an effect on item difficulty in the intended direction (Hypothesis 1a). Moreover, the *change*-manipulation will not affect the effect of person covariates in changed items compared to original items (Hypothesis 1b).

Following the general steps for the *eliminate*-approach, we expected the following: *Eliminate*-items are easier than original items (Hypothesis 2a). Furthermore, by applying the *eliminate*-

manipulation, the effect of person covariates in *eliminate*-items will be changed compared to original items (Hypothesis 2b).

3.4. Method

3.4.1. Sample

Both item manipulations were embedded in a calibration study of the ICT skills test. A sample of $N = 983$ (51% male, 46% female, 3% not specified) was assessed. Participants were between 14 and 17 years ($M = 15.29$, $SD = 0.66$) and from 34 German schools from two federal states in Germany (Baden-Württemberg and Rheinland-Pfalz). Eleven schools belonged to the highest track (Gymnasium), and 23 schools to lower tracks.

3.4.2. Measures of person variables

The initial item pool consisted of 70 items which were implemented in a simulation environment by means of the CBA ItemBuilder (Rölke, 2012). The simulated applications in most items are browsers, e-mails, file managers, text processing software, spread sheet and presentation software (cf. Fig. 1). Items were scored dichotomously. Behavior that could not classified as being definitely right or wrong was treated as neutral and did not count for the final score. For the given example we dealt with this in the following way: Three e-mails (first, third and fifth) should not be forwarded, the fourth e-mail has to be forwarded, and for the second e-mail both solutions are treated as correct. If the test-taker decided for one of the e-mails wrongly, the item was scored as incorrect (0), otherwise as correct (1). For the 70 original items, a one-dimensional Rasch model was fitted using TAM (Kiefer, Robitzsch, & Wu, 2016). Item-infits ranged between 0.87 and 1.11 and item-outfits between 0.67 and 2.18. Two items were excluded from all analyses because of insufficient item-fit. The reliability of the model with 68 items was 0.70. The 42 original items that were manipulated and used for analyses had an average proportion of correct answers of $M = 0.47$ ($SD = 0.26$; $Min = 0.04$, $Max = 0.93$) and were thus from the whole range of difficulties.

As construct-related variables, technical knowledge and the frequency of ICT use were included. A subscale of the Computer Literacy Inventory (INCOBI-R; Richter, Naumann, & Horz, 2010) was used that assesses declarative computer knowledge with 20 multiple-choice items. Scores were computed by a total mean of correct answers ($M = 0.39$, $SD = 0.16$, $\alpha = 0.68$) and z-standardized for data analyses.

To assess ICT use, we asked students to estimate the frequency of seven specific activities in ICT environments in their daily lives. These activities were adapted from the PISA ICT Familiarity questionnaire (OECD, 2013) and assumed to represent such activities that have to be performed also in the test. These are how often they read and write e-mails, search for information for leisure or for school, read texts, create presentations and calculate for mathematics. We used a 4-point likert-scale with response categories “never”, “several times a month”, “several times a week”, and “daily or almost daily”. The variable “ICT use” represents the mean of those seven relevant activities ($\alpha = 0.73$) and was z-standardized.

3.4.3. Item manipulations

To create *change*-items, 40 items were selected from the 70 items. They were selected to be distributed across the five ICT-skills aspects access, manage, integrate, evaluate, and create nearly equally in order to have a good representation of the ICT specific aspects (access, manage and create: eight items; integrate: seven items; evaluate: nine items). Whether items were made easier or harder was very specific to the items. If an item could be assumed to be hard on theoretical grounds, the item was changed to become easier. Correspondingly, if an item could be assumed to be easy on theoretical grounds, it was changed to become harder. From the 40 items, 30 items were intended to become easier and 10 items intended to become harder. Since younger persons struggle with evaluation tasks (Eshet-Alkali & Amichai-Hamburger, 2004), the example item was assumed to be already comparatively hard (the hoax e-mail was sent from a trustworthy person). Thus, we opted for a change that presumably would decrease the item’s difficulty by changing the author in a less trustworthy mailing list. Likewise, a possibility of increasing the item’s difficulty could have been to introduce an even more trustworthy person as the sender of the spam-e-mail (e.g. a supervisor). We applied only one manipulation per item, since this allowed us to use the available testing time to vary rather more items instead of varying one item in two different directions, which is important in the face of a heterogeneous item pool. For a smaller or less heterogeneous item pool, an even more ambitious procedure could include giving some test takers an easier item-variant (e.g. author is a mailing list) and other test-takers the harder item-variant (e.g. author is a supervisor). To create *eliminate*-items, 20 items were selected and stripped of any requirements to apply higher-order ICT skills involving judgement and decision making. These 20 items were equally distributed across the five ICT aspects. Excluding two items led finally to 38 *change*-items (29 easier, 9 harder) and 18 *eliminate*-items for analyses.

3.4.4. Procedures

The assessment consisted of two parts (cf. Table 2), while each part took about one hour. Before the students started with the test, all received a tutorial to become familiar with the simulated environment. Then, students were assigned randomly to the different booklets, and worked in the first part either on original items ($n = 773$) or *change*-items ($n = 210$), but never on both. In the second part of the assessment, *eliminate*-items and questions for technical knowledge and ICT use were administered. From those students who worked in part one on original items, 220 students received *eliminate*-items in the second part of the test. Due to a balanced design of original items in the first part, regarding information tasks, applications, and estimated time intensities (Wenzel et al., 2016). Some of those students ($n = 173$) received in part one of the test already an original version of an *eliminate*-item. Although we minimized this number of overlapping items, this happened on average for three items per person ($M = 2.98$, $SD = 2.09$). As a consequence, the answer on the corresponding *eliminate*-item was not used for analyses, in order to avoid that a second presentation of the same item could have affected the results. Questions regarding ICT use were administered to all students, while a few students ($n = 284$) did not receive the technical knowledge questions.

This design was chosen to ensure at first a well-balanced design for the 70 original items for calibration, by administering as many original items as possible to one student and by balancing items regarding content and time-intensity. We decided to administer change-items parallel to original items in the first part to avoid for motivational reasons that students worked in both parts on demanding and time-intensive ICT skills items. Besides, administering change-items alike eliminate-items in the second part would also have led again to a second presentation of item-variants. This could not be avoided due to strong overlaps in original, change- and eliminate-variants.

3.4.5. Data analyses

Generalized linear mixed models (GLMM; De Boeck et al., 2011; Wilson, De Boeck, & Carstensen, 2008) available in the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014) were used for all hypotheses. With GLMM we refer to a more

Table 2
Design of the study.

Group	Part 1			Part 2	$N = 983$
1		Original Items	B	Eliminate-Items + ICT Use +Technical Knowledge	220
2		Original Items	R	ICT Use + Technical Knowledge	269
3	Tutorial	Original Items	E A	ICT Use	284
4		Change-Items	K	ICT Use + Technical Knowledge	210

general analysis framework allowing for IRT models being explanatory on item side (cf. LLTM; Fischer, 1973) but also doubly explanatory including both item and person covariates (latent regression LLTM), as well as including an error component on item side (LLTM+e; Janssen, Tuerlinckx, Meulders, & DeBoeck, 2003) as random effect (cf. Wilson et al., 2008). In a GLMM, the probability to solve an item correctly is expressed by the logit of the probability P to solve the item correctly, which can be explained by fixed effects, denoted by the Greek letter "β", and random effects, denoted by the Latin letter "b". Equation (1) contains the model that was applied for all analyses. The effect β_0 represents an overall intercept. If also group-specific intercepts β_{0k} are modeled to compare the original to the manipulated items, β_0 refers only to the reference group of original items. To relate each manipulated item to the corresponding original item, the corresponding items were treated as equal but differed in their group membership g, which led also to a group specific random item intercept, b_{0ig}, representing the (residual) item easiness. The random person intercept b_{0p}, represents (residual) person ability. Since students were nested in schools, we also included a random intercept for schools, b_{0s}. A fixed effect β_{0k} was modeled to investigate whether the manipulated items became indeed easier and harder (k) compared to the original items (β_0). For Hypotheses 1b and 2b, additional fixed effects were modeled to investigate whether the manipulated items differ in their relation to the person covariate (v), β_{vk}, from the relation of the original items to the person covariate, β_v.

$$ln\left[\frac{P_{pi}}{1-P_{pi}}\right] = \beta_0 + \sum_{k=1}^{K}\beta_{0k}X_{(p,i)0k} + \beta_v X_{(p,i)v} + \sum_{k=1}^{K}\beta_{vk}X_{(p,i)v}X_{(p,i)0k} + b_{0ig} + b_{0p} + b_{0s} \quad (1)$$

GLMMs include the negative difficulty as item parameter, that is, higher and positive values describe a higher probability of successful task solution and thus easier items. The easiness of an item is represented by the fixed intercept for all items and item-specific deviation from this.

3.5. Results

3.5.1. Change

In line with Hypothesis 1a (Table 3), *change*-manipulations worked in both directions. Items that were intended to become easier were indeed easier than the original items ($\beta = 0.54$, $p < 0.001$) and items that were intended to become harder were indeed harder than the original items ($\beta = -0.90$, $p < 0.001$).

To investigate the influence of construct-related variables (Hypothesis 1b), a model was estimated for the 38 original items and their counterparts, the manipulated *change*-items. The results of Hypothesis 1b (Table 4) indicated that both ICT-related variables are as expected positively related to the probability of success in the original items (technical knowledge: $\beta = 0.29$, $p < 0.001$; ICT use: $\beta = 0.09$, $p = 0.006$). Also in line with the hypothesis, the easier *change*-items did not differ from this relationship (technical knowledge: $\beta = -0.09$, $p = 0.196$; ICT use: $\beta = -0.04$, $p = 0.611$), and the harder items differed only for technical knowledge into the positive direction (technical knowledge: $\beta = 0.21$, $p = 0.037$; ICT use: $\beta = 0.10$, $p = 0.337$), which means that the probability of success in these items was even stronger related to technical knowledge than for the original items.

Results from the *change*-approach support, that the *change*-items were as intended easier or harder, and seemed to measure still the same construct.

3.5.2. Eliminate

Supporting Hypothesis 2a (Table 5), the *eliminate*-items were indeed easier than their original counterparts ($\beta = 1.45$, $p < 0.001$).

To investigate whether manipulations affected the measured construct (Hypothesis 2b; Table 6) the relation of construct-related variables to the probability of success was estimated for original items, again for the 18 original items and their *eliminate*-counterparts.

The results of Hypothesis 2b indicated, that both ICT-related variables were as expected positively related to the probability of success in the original items (technical knowledge: $\beta = 0.25$, $p < 0.001$; ICT use: $\beta = 0.11$, $p = 0.017$). In line with the hypothesis, the relation to ICT use differed for the *eliminate*-items indeed from this relation ($\beta = -0.24$, $p < 0.001$), however, the relation to technical knowledge did not differ for the *eliminate*-items ($\beta = -0.00$,

Table 3
Hypothesis 1a: Probability to solve change-items (38) compared to original items (38).

Parameters			β	*SE*	*z*	*p*
Fixed						
Intercept (Original)			−0.13	0.27	−0.48	0.629
Change Items: Intended easier (29)			0.54	0.15	3.56	<0.001
Change Items: Intended harder (9)			−0.90	0.26	−3.50	<0.001
Random						
Variance (person)		0.38				
Variance (school)		0.16				
Variance (item)	Intercept (Original)	2.48				
	Change	0.55				

Note. Model: value ~ Intended Change + (group | item) + (1 | person) + (1 | school); persons = 973, schools = 34, number of observations (persons x answered items) = 16260.

Table 4
Hypothesis 1b: Estimated effects of the technical knowledge and ICT use. Change-items are compared in their relation to the two variables (Change (easier/harder): Variable) to the relation of the original items to the two variables (Variable (Original)).

			Technical Knowledge				ICT Use			
Parameters			β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Fixed										
Intercept (Original)			−0.12	0.27	−0.44	0.662	−0.12	0.27	−0.45	0.650
Change Items: Intended easier (29)			0.54	0.15	3.56	<0.001	0.56	0.15	3.65	<0.001
Change Items: Intended harder (9)			−0.98	0.25	−3.91	<0.001	−0.89	0.26	−3.47	<0.001
Variable (Original)			0.29	0.04	6.75	<0.001	0.09	0.03	2.72	0.006
Change (easier): Variable			−0.09	0.07	−1.29	0.196	−0.04	0.07	−0.51	0.611
Change (harder): Variable			0.21	0.10	2.08	0.037	0.10	0.10	0.96	0.337
Random										
Variance (person)		0.37					0.38			
Variance (school)		0.11					0.15			
Variance (item)	Intercept (Original)	2.60					2.48			
	Change	0.50					0.54			

Note. Models: value ~ Intended Change * variable + (group | item) + (1 | person) + (1 | school); Model for Technical Knowledge: persons = 681, schools = 34, number of observations (persons x answered items) = 12166; Model for ICT Use: persons = 948, schools = 34, number of observations (persons x answered items) = 15952.

Table 5
Hypothesis 2a: Probability to solve eliminate-items (18) compared to original items (18).

Parameters			β	*SE*	*z*	*p*
Fixed						
Intercept (Original)			−0.24	0.32	−0.76	0.446
Eliminate Items			1.45	0.31	4.72	<0.001
Random						
Variance (person)		0.52				
Variance (school)		0.25				
Variance (item)	Intercept (Original)	1.65				
	Eliminate	1.58				

Note. Model: value ~ group + (group | item) + (1 | person) + (1 | school); persons = 762, schools = 34, number of observations (persons x answered items) = 7944.

Table 6
Hypothesis 2b: Estimated effects of the technical knowledge and ICT use. Eliminate-items are compared in their relation to these variables (Eliminate: Variable) to the relation of the original items to these variables (Variable (Original)).

			Technical Knowledge				ICT Use			
Parameters			β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Fixed										
Intercept (Original)			−0.29	0.31	−0.93	0.355	−0.25	0.32	−0.78	0.437
Eliminate Items			1.51	0.32	4.70	<0.001	1.45	0.31	4.72	<0.001
Variable (Original)			0.25	0.06	4.05	<0.001	0.11	0.05	2.38	0.017
Eliminate: Variable			−0.00	0.08	−0.05	0.961	−0.24	0.07	−3.34	0.001
Random										
Variance (person)		0.58					0.51			
Variance (school)		0.19					0.24			
Variance (item)	Intercept (Original)	1.59					1.65			
	Eliminate	1.73					1.58			

Note. Models: value ~ group * variable + (group | item) + (1 | person) + (1 | school); Model for Technical Knowledge: persons = 479, schools = 34, number of observations (persons x answered items) = 5632; Model for ICT Use: persons = 742, schools = 34, number of observations (persons x answered items) = 7782.

$p = 0.961$).

Results from the *eliminate*-approach support, that *eliminate*-items were as intended easier and seemed to measure a (partly) different construct, since the relation to ICT use changed but not the relation to technical knowledge.

4. Discussion

In the present paper, we introduced two novel approaches to validate test items, *eliminate* and *change*. These approaches allow to relate task characteristics to test scores and can be applied even to heterogeneous items sets, as they are more the rule than the exception in “modern educational assessments” (Baumert, Lüdtke, Trautwein, & Brunner, 2009, p. 166), and also used in the assessment of 21st century skills. Such constructs are often assessed in a contextualized way, which makes the items rather complex. The suggested approaches are particularly useful to investigate whether test scores represent indeed differences in the targeted processes. Using ICT skills as an example, results indicated that *changing* item-specific task characteristics in the items affected item difficulty in the intended direction. These changes did not affect the to-be-measured construct, since the relations to technical knowledge and ICT use were not affected by the manipulation. Only the probability of success in those items that were manipulated to be harder were even stronger explained by technical knowledge compared to the original items. *Eliminating* the targeted skill dimension led to easier items and affected the to-be-measured construct partly, since the relation to ICT use is different for

eliminate-items but not the relation to technical knowledge. In the following, we discuss these results and interpretations especially regarding technical knowledge and what can be gained from these approaches for test development.

4.1. Consequences for test score interpretation of the ICT skills test

How can we interpret the results regarding the targeted test score interpretation? Although the relation to the construct-related variables did as expected not change by applying the *change*-manipulation, items that were manipulated to become harder had an even stronger relationship to technical knowledge (cf. Table 4). This might be because task characteristics requiring already technical knowledge (e.g. knowledge about spam e-mails) are, beside other task characteristics, likely starting points for manipulations. That technical knowledge is even more decisive in items that were manipulated in harder direction does not necessarily speak against the targeted construct interpretation. In the example item for instance (Fig. 1), knowledge about spam is required to identify typical markers of spam and to decide correctly not to forward the hoax e-mail. If for instance a hoax e-mail was sent by a more trustworthy author (e.g. a supervisor instead of a colleague), knowledge about hoax e-mails is likely to be even more decisive for a correct task solution. Test score interpretation would have been rather called into question if these harder items were less related to technical knowledge than the original items. Besides, the relation to ICT use supports that test scores from both *change*-groups can be interpreted in a similar way as the original items, thus, that changing the difficulty of those items did not change the construct.

Against our expectation, the relation to technical knowledge was not affected by the *eliminate*-manipulation. Since we assume that technical knowledge is an integral part of higher-order ICT skills, eliminating higher-order ICT skills should also have affected the relation to technical knowledge. Thus, we expected that when items do not require applying such knowledge, for instance about hoax emails as it is the case in *eliminate*-items, this relation should be affected. What does this mean for the test score interpretation? At first, that we manipulated not those task characteristics in *eliminate*-items that cause the relation to test scores from technical knowledge. Thus, we did either manipulate the wrong task characteristics, or technical knowledge scores represent not, or not only, knowledge that we assumed to be relevant for higher-order skills. That the relation to technical knowledge could be even increased by the *change*-manipulation, supports that the identified task characteristics were somehow related to technical knowledge. This is why we should have a closer look to what test scores from technical knowledge might represent and what we understood by technical knowledge.

From the understanding in our study, technical knowledge plays a double role in the construct we focus on, in ICT skills. At first as integral part of higher-order ICT skills, but also on a lower level as part of basic ICT skills as they are required for navigating (Goldhammer et al., 2013). Finding for instance a forward button (cf. Fig. 1) may require some knowledge about e-mail environments. The scale we used for technical knowledge might possibly not differentiate between technical knowledge that is related to lower and higher-order skills. Thus, even if it is possible to increase the relation to technical knowledge by manipulating knowledge as in the *change*-manipulation, it might be not possible to eliminate this relationship completely because navigation might still require to some extent technical knowledge as it is represented by the scale.

Possibly, we underestimated the role of technical knowledge for merely interacting with ICT environments. However, this does not minor the relevance of the validation approach, but rather implies that technical knowledge as chosen variable was not appropriate. However, the relation to ICT use supports that test scores from *eliminate*-items cannot be interpreted in the same way as the original items, which is strongly supported by the even negative relation to ICT use for the *eliminate*-items. Thus, we changed the construct at least partly with the *eliminate*-manipulation.

Further item manipulations could help to investigate the role of technical knowledge. One manipulation could contain, for instance, to keep only higher-order processes in the item by eliminating the navigation from the item. This could be reached by presenting for instance screenshots of the items and to compare then the relation of technical knowledge to the probability of success in such items to the relation of technical knowledge to the probability of success for original items. If the relationship changes, technical knowledge is indeed required for navigation. Taken together, we learned that the entanglement of different levels of skills involved in CBA items is not trivial and that specific attention should be paid to validity of test scores assessed with complex items as they are used for instance in educational assessments.

4.2. Deeper analyses and implications for test developers

The suggested approaches can provide valuable and additional information regarding the single items and task characteristics for test developers. Although the used method allows investigating at first only the average change of item difficulties due to the applied manipulation, deeper analyses can be conducted.

Firstly, it can be analyzed whether the changes were differently effective in different items by referring to the variance in items above the average effect. This can be reached by comparing a model with item-specific adaptation to the *change*-effect (with random effect) to a model without item-specific adaptation (no random effect). Such analyses are especially useful if there was no average change in item difficulty, to investigate for instance whether only a few items did not change, or whether all changes were too small. Item-specific adaptations to the average intercepts can indicate which items changed most or less, or even in the wrong direction. If an item did not change, this can be for instance because the manipulated task characteristic was not at all used by the test-takers as assumed (e.g. evaluation processes were not performed at all), because the change was not effective and did not affect the evaluation process, or because the original item was already very easy or hard for the test-takers.

Secondly, it can be helpful to group items regarding task characteristics, for instance, items that require similar evaluation processes, if the number of items per task characteristic is sufficient and the selected task characteristics for the grouping are meaningful. This allows analyzing whether indeed all groups of task characteristics affected item difficulty. These deeper analyses can help reconsidering theoretical assumptions and indicators (cf. Kane, 2013, p. 40).

4.3. Conclusion

Using experimental strategies for test score validation, if successful, can support the plausibility of test score interpretation, because the targeted test score interpretation is challenged. Although the process of validation depends on the test and the construct, the *eliminate*- and *change*-approaches provide a general strategy for validation that can be transferred to other constructs and contexts. This is especially the case in the area of educational measurement, where broad constructs are used. These constructs are often assessed by means of computers, allowing the simulation of authentic settings. This may lead the same time to heterogeneous item sets, where current validation approaches cannot be

applied to. The two suggested strategies combine experimental techniques with the recent concept of validation. They provide a concrete and systematic approach for implementing the modern understanding of validity. For this reason they can be regarded as a valuable tool assuring a theory-based operationalization of constructs through test items.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research [grant numbers: 01LSA010, 01LSA010A, 01LSA010B]. We would like to thank Olga Kunina-Habenicht for her comments on an earlier version of the manuscript.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington: AERA, APA, NCME.

Wenzel, S. F. C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F., Naumann, J., & Horz, H. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) (CavE-ICT) [Computer-based, adaptive and behavior-related assessment of information and communication-related competencies (ICT skills)]. In , *Forschung in Ankopplung an Large-scale AssessmentsBMBF (Hrsg.)* (pp. 161–180). Bonn, Berlin: BMBF.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using eigen and S4*. R package version 1.1-7 http://CRAN.R-project.org/package=lme4.

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*, 165–176.

Binkley, M., Erstad, O., Herman, J., Raizen, R., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht: Springer.

Caissie, A. F., Vigneau, F., & Bors, D. A. (2009). What does the Mental Rotation Test measure? An analysis of item difficulty and item characteristics. *Open Psychology Journal, 2*, 94–102.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1–28.

van Deursen, A. J., & van Dijk, J. A. (2009). Using the Internet: Skill related problems in users' online behavior. *Interacting with Computers, 21*, 393–402. http://dx.doi.org/10.1016/j.intcom.2009.06.005.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Eshet-Alkalai, Y., & Chajut, E. (2010). You can teach old dogs new tricks: The factors that affect changes over time in digital literacy. *Journal of Information Technology Education, 9*, 173–181.

Eshet-Alkali, Y., & Amichai-Hamburger, Y. (2004). Experiments in digital literacy. *Cyber Psychology & Behavior, 7*, 421–429.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374. http://dx.doi.org/10.1016/0001-6918(73)90003-6.

Fraillon, J., & Ainley, J. (2010). *The IEA international study of computer and information literacy (ICILS)*. Retrieved from http://www.researchgate.net/profile/John_Ainley/publication/268297993_The_IEA_International_Study_of_Computer_and_Information_Literacy_%28ICILS%29/links/54eba4330cf2082851be49a9.pdf.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age*. Springer-Verlag GmbH.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*, 263–275.

Greiff, S., Krkovic, K., & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56*, 83–103.

Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior, 55*, 486–500. http://dx.doi.org/10.1016/j.chb.2015.09.042.

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten [Using the prediction of item difficulties for construct validation and model-based proficiency scaling]. *Psychologische Rundschau, 63*, 43–49. http://dx.doi.org/10.1026/0033-3042/a000109.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*, 369–380.

International ICT Literacy Panel. (2002). *Digital transformation: A framework for ICT literacy*. NJ: Princeton. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2002/cjik.

Janssen, R., Tuerlinckx, F., Meulders, M., & DeBoeck, P. (2003). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285–306. http://dx.doi.org/10.2307/1165207.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules*. R package version 1.16-0 http://CRAN.R-project.org/package=TAM.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15–21.

OECD. (2011). *PISA 2009 Results: Students on Line: Digital technologies and performance (volume VI)*. http://dx.doi.org/10.1787/9789264112995-en.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. OECD Publishing. http://dx.doi.org/10.1787/9789264128859-en.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en.

OECD. (2014). *PISA 2012 results: What students know and can do – student performance in mathematics, reading and science (volume I, revised edition, February 2014)*. PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en.

Pfaff, Y., & Goldhammer, F. (2011, September). Measuring individual differences in ICT literacy: Evaluating online information. In *Talk presented at the 14th biennial conference of the European Association for research on learning and instruction (EARLI), Exeter, United Kingdom*.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Richter, T., Naumann, J., & Horz, H. (2010). Das Inventar zur Computerbildung (revidierte Fassung). *Zeitschrift für Pädagogische Psychologie, 24*, 23–37.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53*, 145–161. http://dx.doi.org/10.1002/asi.10017.

Rölke, H. (2012). The item builder: A graphical authoring system for complex item development. In T. Bastiaens, & G. Marks (Eds.), *Proceedings of world conference on e-learning in corporate, government, healthcare, and higher education* (pp. 344–353). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/41614.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior, 65*, 100–106.

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal of Psychological Assessment, 18*, 190–203. http://dx.doi.org/10.1027//1015-5759.18.3.190.

Whittaker, S., & Sidner, C. (1996, April). Email overload: Exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 276–283)*. ACM.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Göttingen: Hogrefe.

Anhang D

Erklärung zur Promotionsordnung

Ich erkläre hiermit, dass mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fachbereiche der Goethe-Universität Frankfurt am Main bekannt ist.

Frankfurt am Main, den 05.10.2017

Lena Engelhardt

Eidesstattliche Versicherung

Ich erkläre hiermit, dass ich die vorgelegte Dissertation mit dem Titel *Fertigkeiten für die Lösung von kognitiven ICT-Aufgaben - Entwicklung und empirische Erprobung eines Erhebungs- und Validierungskonzepts* selbständig angefertigt und nur die in der Dissertation angegebenen Hilfsmittel benutzt habe. Ich versichere, dass alle Entlehnungen aus anderen Schriften mit Angabe der betreffenden Schrift gekennzeichnet sind und ich die Grundsätze der guten wissenschaftlichen Praxis beachtet habe.

Frankfurt am Main, den 05.10.2017

Lena Engelhardt

Erklärungen über frühere Promotionsversuche

Hiermit erkläre ich, dass keine früheren Promotionsversuche vorliegen.

Frankfurt am Main, den 05.10.2017

Lena Engelhardt

Anhang E

Stellungnahme zu den Kriterien einer publikationsbasierten Dissertation

Kriterien für kumulative Dissertationen im Fachbereich Psychologie und Sportwissenschaften, Goethe Universität Frankfurt (nach Beschluss im Fachbereichsrat gültig ab 11.06.2015)

(1) Die kumulative Dissertation soll in der Regel 3 Schriften umfassen, die aus den letzten 5 Jahren stammen sollen.
Erklärung: Die Dissertation umfasst eine Schrift aus dem Jahr 2017 sowie zwei derzeit in Fachzeitschriften eingereichte Manuskripte.

1. Schrift:
   Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Horz, H., Hartig, K., & Wenzel, S. F. C. (*submitted, International Journal of Testing*). A framework for the performance-based testing of ICT skills.
2. Schrift:
   Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Wenzel, S. F. C., Hartig, K., & Horz, H. (*submitted, European Journal of Psychological Assessment*). Convergent evidence for validity of a performance-based ICT skills test.
3. Schrift:
   Engelhardt, L., Goldhammer, F., Naumann, J., & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. *Computers in Human Behavior, 76*, 683-692. doi: 10.1016/j.chb.2017.02.020

(2) Die Schriften sollen im Wesentlichen einem zusammenhängenden Forschungsprogramm entstammen. Die jeweils verfolgten Forschungsfragen sollen sich sinnvoll zueinander in Beziehung setzen lassen.
Erklärung: Die Schriften basieren alle auf Forschungsdaten die im Rahmen des Verbundprojekts CavE-ICT-PISA erhoben wurden. Darüber hinaus thematisieren alle drei Schriften die simulationsbasierte Erhebung von ICT-Skills. Während die erste Arbeit im Wesentlichen die Rahmenkonzeption für die Erstellung der Items beschreibt,

beschäftigen sich die Arbeiten zwei und drei mit der Validierung der Testwerteinterpretation.

(3) Der Kandidat oder die Kandidatin soll bei 2 Publikationen Erstautor/Erstautorin sein, bei einer weiteren Publikation kann er/sie Koautor/Koautorin sein. Eine geteilte Erstautorenschaft wird für jeden der Erstautoren anteilig gewichtet (bei 2 Erstautoren eine 1/2 Erstautorenschaft, bei 3 eine 1/3 Erstautorenschaft usw.).
Erklärung: Alle drei Schriften sind in Erstautorenschaft der Kandidatin verfasst.

(4) Die drei Schriften sollen zur Veröffentlichung zumindest eingereicht sein. Der aktuelle Status ist detailliert darzulegen (Publikationsorgan und Status wie eingereicht, in revision, conditional accept usw.).

(5) Mindestens 2 der 3 Schriften müssen in guten oder sehr guten, in der Regel englischsprachigen, Zeitschriften mit Peer-Review eingereicht sein.

(6) Eine der 3 Schriften kann als Publikation in einem einschlägigen Lehrbuch, Enzyklopädieband oder einem anderen für das jeweilige Fach bedeutsamen Publikationsorgan, jeweils mit Peer-Review, eingereicht oder veröffentlicht sein.
Erklärung zu 4-6: Die dritte Schrift ist in der englischsprachigen Zeitschrift Computers in Human Behavior veröffentlicht (Impact Factor 2016: 3.435). Die erste Schrift wurde in der Zeitschrift International Journal of Testing eingereicht, die zweite Schrift in der Zeitschrift European Journal of Psychological Assessment. Alle drei Zeitschriften sind englischsprachig und arbeiten mit Peer-Review Verfahren.

(7) Die als Dissertation vorgelegte Abhandlung soll über die zusammengestellten Publikationen hinaus einen zusätzlichen Text enthalten, in welchem eine kritische Einordnung der eigenen Publikationen aus einer übergeordneten Perspektive heraus vorgenommen wird. Dieser Text sollte einen Umfang von ca. 30 Seiten haben. Es sollen die Fragestellungen theoretisch entwickelt werden, die empirischen Arbeiten und ihre Ergebnisse so dargestellt werden, dass sie auch ohne Lesen der Einzelarbeiten nachvollziehbar sind und es soll eine Gesamtdiskussion enthalten, die die Fragestellungen beantwortet und den Erkenntnisgewinn der Arbeit herausstellt.

Erklärung: Die Dissertation enthält über die drei Schriften hinaus einen deutschsprachigen Text, der die Fragestellungen entwickelt, die Ergebnisse aller drei Arbeiten darstellt und diese aus übergeordneter Perspektive diskutiert.

(8) Die Dissertation muss eine Erklärung enthalten, in der die Eigenleistung des Kandidaten/der Kandidatin dargestellt wird. Insbesondere bei Schriften mit Koautoren, aber auch bei in Einzelautorenschaft entstandenen Schriften, die oft auch im Rahmen von Abteilungsprojekten, Drittmittelprojekten, Projektverbünden usw. entstanden sind, soll dargelegt werden, welchen Anteil die Kandidaten an Entwicklung der Fragestellung, Design, Durchführung, Auswertung der empirischen Studie(n) und an dem Abfassen der einzelnen Beiträge hatten. Diese Erklärung ist von Betreuer und/oder Koautoren zu bestätigen.
Erklärung: Die Erklärung zur Eigenleistung ist in Anhang F enthalten.

(9) In besonders begründeten Fällen kann von diesen Richtlinien abgewichen werden.

(10) Bei den vorgeschlagenen Kriterien handelt es sich um Empfehlungen. Es wird explizit darauf hingewiesen, dass natürlich nach wie vor die jeweilige Promotionsordnung, die Beschlüsse des Promotionsausschusses und die von den Gutachtern erstellten Gutachten entscheidend für das Verfahren sind.

Anmerkung: Satz (8) gilt auch für Dissertationen, die als Monographie vorgelegt werden.

Frankfurt am Main, den 05.10.2017

Lena Engelhardt

Anhang F

Erklärung über die Eigenleistung der Kandidatin

Die vorliegende Dissertation *Fertigkeiten für die Lösung von kognitiven ICT-Aufgaben - Entwicklung und empirische Erprobung eines Erhebungs- und Validierungskonzepts* beinhaltet drei Manuskripte, die alle auf Daten aus dem Verbundprojekt CavE-ICT PISA basieren. Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Projekt wurde von Prof. Dr. Holger Horz (Goethe-Universität Frankfurt), Prof. Dr. Andreas Frey (Friedrich-Schiller-Universität Jena) und Prof. Dr. Frank Goldhammer (DIPF, Frankfurt) eingeworben. Im Rahmen des Projekts (Laufzeit 4/2012 - 3/2015) wurde ein Test zur computergestützten, adaptiven und verhaltensnahen Erfassung Informations- und Kommunikationstechnologiebezogener Fähigkeiten entwickelt und erprobt. Die Kandidatin arbeitete in dem Projekt am DIPF als wissenschaftliche Mitarbeiterin unter der gemeinsamen Projektleitung von Prof. Dr. Frank Goldhammer und Prof. Dr. Johannes Naumann. Die Kandidatin hat maßgeblich und eigenverantwortlich an der Entwicklung und Implementierung der Items, bei der Erstellung des Testsystems sowie an der technischen Auslieferung mitgewirkt. Die Studie wurde gemeinsam mit dem Data Processing and Research Center (IEA DPC) durchgeführt, wobei die Kandidatin an der Studiendurchführung beteiligt war (u.a. Testleiterschulung, technischer Support in der Erhebungsphase). Die Daten wurden am DIPF ausgelesen und standen seit Herbst 2014 zur Verfügung.

Schrift 1: Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Horz, H., Hartig, K., & Wenzel, S. F. C. (*submitted, International Journal of Testing*). A framework for the performance-based testing of ICT skills.

Die Projektgruppe, die bei dieser Schrift in Koautorenschaft beteiligt ist, entwickelte zusammen das Assessmentframework, das als Grundlage für die Itementwicklung gedient hat. Dieses Manuskript basiert auf Teilen dieses Assessmentframeworks. Die beiden konzeptuellen Beiträge des Manuskripts, die theoretische Untermauerung der Rahmenkonzeption sowie die verhaltensbasierte Implementation wurden von der Kandidatin eigenständig entwickelt und ausgeführt und mit der Projektgruppe diskutiert. Die Ideen für die empirischen Analysen hat die Kandidatin selbständig entwickelt und umgesetzt. Die Kandidatin verfasste das Manuskript selbständig, wobei die Koautoren dieses inhaltlich kommentierten und sprachlich überarbeiteten.

Schrift 2: Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Wenzel, S. F. C., Hartig, K., & Horz, H. (*submitted, European Journal of Psychological Assessment*). Convergent evidence for validity of a performance-based ICT skills test.

Die Auswahl der Validierungsinstrumente wurde gemeinsam in der Projektgruppe vorgenommen, die an dem Manuskript in Koautorenschaft beteiligt ist. Die Kandidatin erarbeitete mit dem Projektteam am DIPF die Fragestellung und diskutierte sie mit den Koautoren. Die Computerisierung der Validierungsinstrumente wurde von der Kandidatin geplant und organisiert. Das Validierungsdesign wurde von dem Projektteam am Standort DIPF, d.h. der Kandidatin und den Projektleitern, entwickelt. Die statistischen Analysen wurden von der Kandidatin eigenständig geplant und durchgeführt. Das Manuskript wurde federführend von der Kandidatin verfasst.

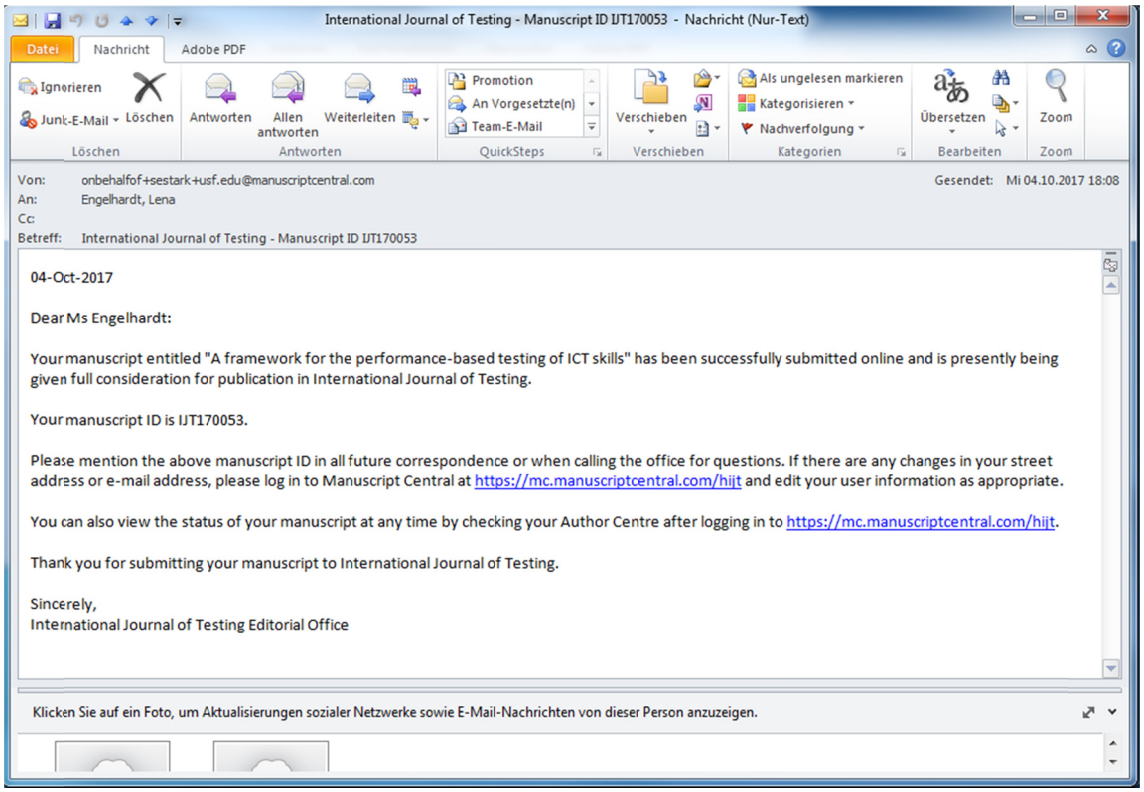
Schrift 3:
Engelhardt, L., Goldhammer, F., Naumann, J., & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. *Computers in Human Behavior, 76*, 683-692. doi: 10.1016/j.chb.2017.02.020

Die Entwicklung der Itemmanipulationen war Gegenstand des Teilprojekts zur Bearbeitung der Validierungsfragestellungen am Standort DIPF. Die Itemvarianten wurden von der Kandidatin unter Beratung von Prof. Dr. Frank Goldhammer und Prof. Dr. Johannes Naumann geplant und umgesetzt. Die Entwicklung sowie praktische Umsetzung der Itemvarianten in der Autorensoftware CBA ItemBuilder wurden von der Kandidatin durchgeführt. Die Hypothesen und statistischen Analysen wurden von der Kandidatin entwickelt und durchgeführt und mit den Koautoren diskutiert. Die Schrift wurde von der Kandidatin verfasst und von den Koautoren inhaltlich kommentiert und sprachlich überarbeitet.

Prof. Dr. Frank Goldhammer	Lena Engelhardt
(Betreuer der Dissertation)	(Verfasserin der Dissertation)

Anhang G

Bestätigung der Einreichungen



International Journal of Testing - Manuscript ID IJT170053 - Nachricht (Nur-Text)

Datei | Nachricht | Adobe PDF

Ignorieren | Junk-E-Mail | Löschen | Antworten | Allen antworten | Weiterleiten | Promotion | An Vorgesetzte(n) | Team-E-Mail | Verschieben | Als ungelesen markieren | Kategorisieren | Nachverfolgung | Übersetzen | Zoom

Von: onbehalfof+sestark+usf.edu@manuscriptcentral.com
An: Engelhardt, Lena
Cc:
Betreff: International Journal of Testing - Manuscript ID IJT170053
Gesendet: Mi 04.10.2017 18:08

04-Oct-2017

Dear Ms Engelhardt:

Your manuscript entitled "A framework for the performance-based testing of ICT skills" has been successfully submitted online and is presently being given full consideration for publication in International Journal of Testing.

Your manuscript ID is IJT170053.

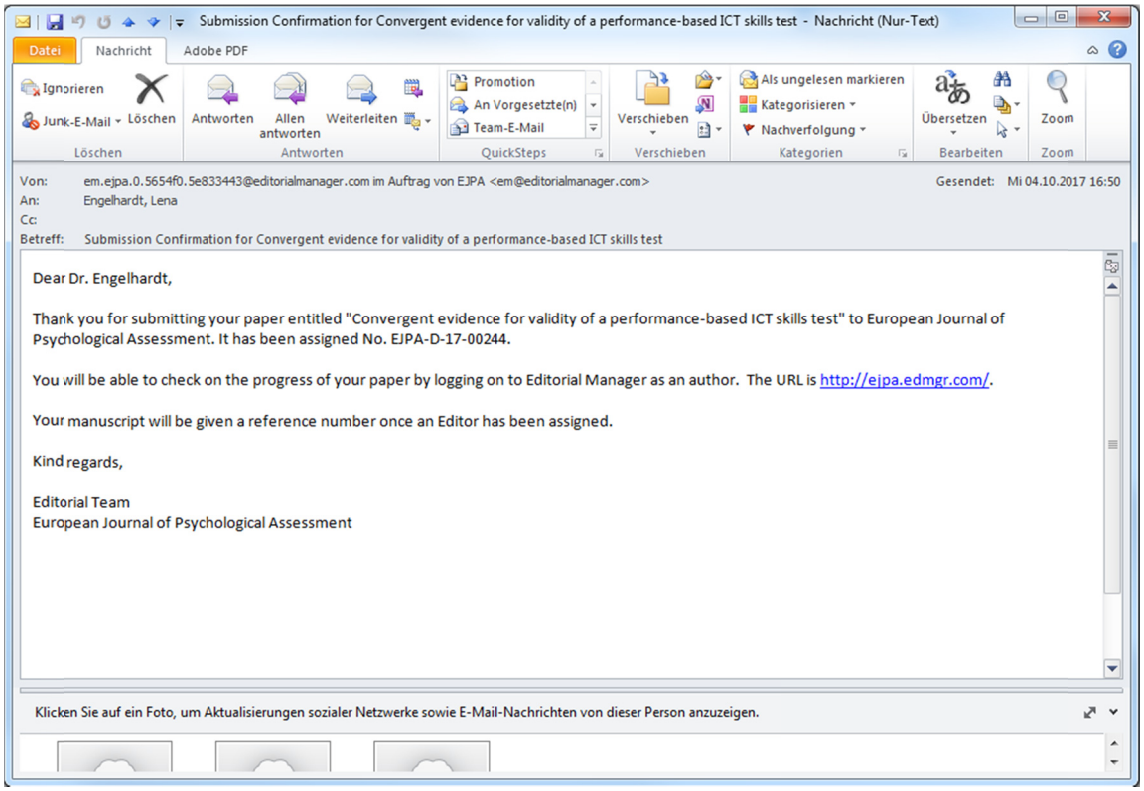
Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to Manuscript Central at https://mc.manuscriptcentral.com/hijt and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Centre after logging in to https://mc.manuscriptcentral.com/hijt.

Thank you for submitting your manuscript to International Journal of Testing.

Sincerely,
International Journal of Testing Editorial Office

Klicken Sie auf ein Foto, um Aktualisierungen sozialer Netzwerke sowie E-Mail-Nachrichten von dieser Person anzuzeigen.



Submission Confirmation for Convergent evidence for validity of a performance-based ICT skills test - Nachricht (Nur-Text)

Datei | Nachricht | Adobe PDF

Ignorieren | Junk-E-Mail | Löschen | Antworten | Allen antworten | Weiterleiten | Promotion | An Vorgesetzte(n) | Team-E-Mail | Verschieben | Als ungelesen markieren | Kategorisieren | Nachverfolgung | Übersetzen | Zoom

Von: em.ejpa.0.5654f0.5e833443@editorialmanager.com im Auftrag von EJPA <em@editorialmanager.com>
An: Engelhardt, Lena
Cc:
Betreff: Submission Confirmation for Convergent evidence for validity of a performance-based ICT skills test
Gesendet: Mi 04.10.2017 16:50

Dear Dr. Engelhardt,

Thank you for submitting your paper entitled "Convergent evidence for validity of a performance-based ICT skills test" to European Journal of Psychological Assessment. It has been assigned No. EJPA-D-17-00244.

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author. The URL is http://ejpa.edmgr.com/.

Your manuscript will be given a reference number once an Editor has been assigned.

Kind regards,

Editorial Team
European Journal of Psychological Assessment

Klicken Sie auf ein Foto, um Aktualisierungen sozialer Netzwerke sowie E-Mail-Nachrichten von dieser Person anzuzeigen.