G Methoden und Technologien des Assessments

Research

G.1 Itempool-Management mit Microsoft Excel: Eine UX-Studie.

Marios Karapanos[1], Andreas Thor[2], Heinz-Werner Wollersheim[1]
[1] Universität Leipzig, Erziehungswissenschaftliche Fakultät
[2] HTWK Leipzig, Fakultät Digitale Transformation

1 Einführung

Elektronische Prüfungen (E-Assessments) mit standardisierten Aufgabenformaten sind in teilnehmerstarken Prüfungssituationen ein besonders effizientes Testverfahren (Michel, Goertz, Radomski, Fritsch, & Baschour, 2015; Pengel, Hawlitschek, & Karapanos, 2019). Gleichzeitig entstehen mit ihrem Einsatz hohe Aufwände bei der Verwaltung der notwendigen Aufgabensammlungen (Itempool-Management). Die gegenwärtig eingesetzten Learning-Management-Systeme (LMS) unterstützen die in diesem Zusammenhang anfallenden Arbeitsaufgaben oft nur unzureichend, obwohl effizientere Methoden technisch machbar erscheinen. Zwar verfügen LMS wie ILIAS oder OPAL grundlegend über alle notwendigen Funktionen zur Erstellung, Bearbeitung und für den Austausch von Testitems. Die browserbasierten grafischen Schnittstellen erfordern allerdings ein hohes Maß an ‚Klickarbeit', sind wegen der technisch bedingten Wartezeiten zwischen Eingabe und Systemantwort bei synchronen Webdiensten nicht immer zeiteffizient und erscheinen damit für den Aufbau und die Pflege großer Itempools wenig geeignet. Die Plattformen offenbaren darüber hinaus Schwächen bei der Erfassung und Bearbeitung wichtiger inhaltsbezogener Metadaten wie der Anforderungsstufe, zugeordneter Learning Outcomes oder der thematischen Verortung eines Items innerhalb einer Wissensdomäne. Auch die vorhandenen Import-/Exportschnittstellen für die Offline-Bearbeitung stellen keine zufriedenstellende Lösung des Problems dar, weil sie wegen des in der Regel genutzten XML-Dateiformats für technische Laien weitgehend gebrauchsuntauglich sind.

Im vorliegenden Beitrag wird ein Interaktionskonzept für das Itempool-Management auf Basis von Microsoft Excel vorgestellt, das die benannten Schwachstellen adressiert und eine Alternative zu bestehenden Lösungen anbietet. Seine praktische Eignung wird anhand von Ergebnissen eines vergleichenden Nutzertests überprüft und diskutiert.

2 Theoretischer Bezugsrahmen

Obschon das Benutzererlebnis (User Experience, kurz UX) und insbesondere Gebrauchstauglichkeit (Usability) als dessen Teilkomponente wichtige Qualitätsmerkale interaktiver Systeme darstellen, werden darauf abzielende Gestaltungsgrundsätze und -leitlinien in Softwaresystemen zum Itempool-Management bislang nur unzureichend berücksichtigt. Als Kernmerkmale nutzerorientierter Gestaltungsprozesse (User Centered Design) gelten die frühzeitige Berücksichtigung des Nutzers und seiner Aufgaben, die empirische Prüfung von Designentwürfen in Nutzertests und ein iterativer Designprozess (Gould & Lewis, 1985). Usability kann definiert werden als „das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um festgelegte Ziele effektiv, effizient und zufriedenstellend zu erreichen“ (DIN EN ISO 9241-11, 1998, S. 4). User Experience geht in der Definition über die instrumentelle Eignung des Systems hinaus und umfasst auch „sämtliche Emotionen, Vorstellungen, Vorlieben, Wahrnehmungen, physiologischen [sic] und psychologischen [sic] Reaktionen, Verhaltensweisen und Leistungen, die sich vor, während und nach der Nutzung ergeben“ (DIN EN ISO 9241-210, 2010, S. 7). Im von Hassenzahl (2007) entwickelten UX-Modell werden zwei unabhängige UX-Dimensionen unterschieden, die pragmatische Qualität und die hedonische Qualität. Während die pragmatische Qualität weitgehend der Usability entspricht, ergibt sich die hedonische Qualität aus der Fähigkeit des Systems, psychogene Grundbedürfnisse wie Autonomie, Stimulation oder auch Sicherheit zu befriedigen (Diefenbach & Hassenzahl, 2017; Sheldon, Elliot, Kim, & Kasser, 2001). Aus dem Zusammenspiel beider Qualitäten resultiert schließlich die Attraktivität eines Systems, wobei die Bedeutung der beiden Qualitäten jeweils variieren kann. Während bspw. bei einem Geldautomaten der Fokus typischerweise auf der pragmatischen Qualität liegt, gewinnt vor allem bei Konsumprodukten wie Smartphones und den darauf installierten Apps die hedonische Qualität stärker an Bedeutung.

3 Itempool-Management mittels Excel-Addin

Ausgehend von einer Analyse bestehender Lern- und Prüfungsplattformen und typischer Aufgaben bei der Erstellung und Pflege großer Itempools wurde ein Interaktionskonzept auf Basis von Microsoft Excel entwickelt und prototypisch an das E-Assessment-Literacy-Tool EAs.LiT (Thor, Pengel, & Wollersheim, 2017) angebunden. Durch diese Kombination lassen sich Funktionalität und Bedienkomfort einer etablierten Tabellenkalkulationssoftware mit der Flexibilität eines webbasierten Datenbanksystems verzahnen und für das Management großer Itempools nutzbar machen. Gängige Workflow-Elemente wie Copy & Paste, Search & Replace oder das Arbeiten mit Formeln ermöglichen das effiziente Erstellen und Editieren mehrerer Items mit gleicher Problemvignette oder gleichen bzw. ähnlichen Antwortoptionen.

Gleichzeitig ermöglicht die übersichtliche tabellarische Darstellung eine Homogenisierung des Itempools z. B. in Bezug auf Schreibweisen, Formulierungen, Layout oder Punkteverteilung über verschiedene Items hinweg. Eine Erweiterung um zusätzliche Itemattribute ist dabei jederzeit durch das Hinzufügen von Tabellenspalten problemlos möglich, sodass auf zukünftige Anforderungen flexibel reagiert werden kann. Die technische Implementation wird dabei als Excel-Addin realisiert, welches über Webservices mit dem zentralen EAs.LiT-System kommuniziert. So kann auf händische Up- und Downloads von Austauschdateien verzichtet und ein insgesamt – so die Zielstellung – hohes Maß an Gebrauchstauglichkeit und eine insgesamt bessere User Experience erzielt werden. Zwar lag der Fokus bei der Entwicklung dieses Interaktionskonzepts primär auf mehr Effektivität und Effizienz und damit einer verbesserten pragmatischen Qualität. Mit dem Einsatz eines vertrauten und erfolgreichen Softwaresystems, wie Excel es darstellt, erscheint aber zudem auch eine Steigerung der hedonischen Qualität zumindest möglich.

4 Methode

Zur Überprüfung des neuen Interaktionskonzepts wurden Nutzertests mit 15 Testpersonen (11 davon weiblich) durchgeführt. Zwar empfiehlt bspw. Nielsen (2012) wenigstens 20 Testpersonen für quantitativ ausgerichtete Nutzertest. Kommt jedoch ein Untersuchungsdesign mit Messwiederholung zum Einsatz und werden zudem große Unterschiede zwischen den getesteten Systemen erwartet, kann die Zahl der Testpersonen auch reduziert werden, weil sich beide Faktoren (Design mit Messwiederholung und hohe Effektstärken) vorteilhaft auf die Teststärke auswirken. Alle Testpersonen wurden aus dem wissenschaftlichen Personal des Lehrstuhls für Allgemeine Pädagogik der Universität Leipzig rekrutiert. Das Durchschnittsalter betrug 31.9 Jahre ($SD = 7.83$). Für den Test wurden drei typische Aufgaben ausgewählt, die im Rahmen des Itempool-Managements regelmäßig zu bewältigen sind:

1. das Anlegen eines neuen Items,
2. das Editieren bestimmter Itemattribute bei einem bestehenden Item und
3. das Editieren des identischen Itemattributs bei mehreren bestehenden Items.

Als Vergleichssysteme kamen die LMS ILIAS und OPAL zum Einsatz. Die Testpersonen bearbeiteten die Aufgaben mit Excel und den zwei Vergleichssystemen (Design mit Messwiederholung) ohne spezielle Einweisung ins jeweilige System. Um Reihenfolgeeffekte zu vermeiden, kam ein balancierter Versuchsplan zum Einsatz. Jedes System war also jeweils fünfmal das erste, das zweite und das dritte System im Testplan. Zur Überprüfung auf statistische Signifikanz werden einfaktorielle ANOVA mit Messwiederholung bzw. im nicht-parametrischen Fall der Friedman-Test angewendet. Post-hoc-Tests werden nach der Bonferroni-Holm-Methode für multiples Testen korrigiert. Als Effektmaß für paarweise Vergleiche wird Cohens *d* angegeben (Cohen, 1988).

Nach Bearbeitung der drei Aufgaben an einem System wurden den Testpersonen jeweils eine deutsche Fassung der System Usability Scale (SUS, Brooke, 1996; Rauer, 2011) und eine 11-Item-Kurzfassung des User Experience Questionnaire (UEQ, Alberola et al., 2017) vorgelegt. Die Kurzfassung des UEQ erfasst das subjektive Erleben in den Dimensionen pragmatische (PQ) und hedonische Qualität (HQ) und enthält eine zusätzliche globale Attraktivitätsskala (ATT). Der Fokus der SUS liegt ebenfalls auf der pragmatischen Qualität. Wegen der großen Zahl verfügbarer Referenzstudien eignet sie sich gut für Benchmarkings und wurde deswegen zusätzlich integriert.

Anhand von Bildschirmaufzeichnungen jeder Testsitzung wurden im Nachgang Erfolgsrate und Bearbeitungszeit je Aufgabe für jedes System ermittelt. Die Erfolgsrate gilt als Indikator für die Effektivität eines Systems, die Bearbeitungszeit für dessen Effizienz. Die Bearbeitungszeit wird hier definiert als die Zeit zwischen dem Beginn einer Aufgabe und dem erfolgreichen oder erfolglosen Abschluss durch die Testperson (Sauro & Lewis, 2016). Da sich Vorerfahrung mit einem System in der Regel positiv auf die Aufgabenbewältigung auswirkt, wird diese für jedes System mittels dreistufiger Rangskala (1 = keine / 2 = wenig / 3 = viel) miterfasst.

5 Ergebnisse

Mit Ausnahme der Skala zur hedonischen Qualität aus dem UEQ zeigen alle Skalen eine zufriedenstellende bis sehr gute interne Konsistenz. Hypothesentests auf Basis der HQ-Skala sind damit nur unter Vorbehalt zu interpretieren.

Tabelle 1: Deskriptive Statistik zu SUS und UEQ

	SUS	ATT	HQ	PQ
Excel				
M (*SD*)	65.8 (19.6)	4.52 (1.20)	3.98 (0.94)	4.68 (1.28)
Cronbachs α	.85	.90	.68	.75
ILIAS				
M (*SD*)	62.5 (16.4)	4.30 (0.97)	4.29 (0.91)	4.45 (0.95)
Cronbachs α	.85	.88	.67	.68
OPAL				
M (*SD*)	43.5 (19.9)	3.85 (1.12)	4.07 (0.90)	3.42 (1.21)
Cronbachs α	.89	.91	.47	.91

Anmerkung. SUS = System Usability Scale, ATT = Attraktivität (UEQ), HQ = Hedonische Qualität (UEQ), PQ = Pragmatische Qualität (UEQ)

Die Testpersonen verfügten über ein unterschiedliches Maß an Erfahrung mit den getesten Systemen (Friedman-Test: $\chi^2 = 19.7$, $p < .001$).Post-hoc-Tests weisen auf signifikante Unterschiede zwischen Excel und den beiden LMS ILIAS und OPAL hin (Conover-Test: $p_{Holm} < .050$). Zwischen ILIAS und OPAL besteht hingegen kein Unterschied (Conover-Test: $p_{Holm} = .125$). Die meiste Erfahrung besaßen die Testpersonen mit Excel ($M = 2.67$, $SD = 0.49$). Mit ILIAS ($M = 1.67$, $SD = 0.21$) und OPAL ($M = 1.20$, $SD = 0.56$) waren sie deutlich weniger vertraut.

Tabelle 2: Erfolgsraten und Bearbeitungszeiten

	Aufgabe 1	Aufgabe 2	Aufgabe 3	gemittelt
Excel Erfolgsrate Bearbeitungszeit in s	53% 150 (45)	93% 104 (68)	73% 100 (71)	73% 353 (156)
ILIAS Erfolgsrate Bearbeitungszeit in s	73% 174 (65)	40% 179 (41)	80% 119 (82)	64% 472 (157)
OPAL Erfolgsrate Bearbeitungszeit in s	20% 180 (47)	80% 126 (42)	53% 156 (55)	51% 461 (101)

Die erfassten UX-Maße korrelieren theoriekonform. Der stärkste Zusammenhang besteht zwischen SUS und PQ-Skala des UEQ, die beide die pragmatische Qualität erfassen. Die Bearbeitungszeit korreliert stark negativ mit der Erfolgsrate, aber mit keinem der Fragebogeninstrumente (siehe Tabelle 3).

5.1 Subjektive Bewertung

Zwischen den getesteten Systemen bestehen keine signifikanten Unterschiede in der hedonischen Qualität ($F(2, 28) = 0.81$, $p = .457$) und der globalen Attraktivität ($F(2, 28) = 1.29$, $p = .292$). Unterschiede in der pragmatischen Qualität zeigen sich aber sowohl in der Messung durch den UEQ ($F(2, 28) = 4.53$, $p = .020$) als auch durch die SUS ($F(2, 28) = 5.76$, $p = .008$). Post-hoc-Tests weisen auf signifikante Unterschiede zwischen Excel und OPAL (UEQ: $t = 2.827$, $p_{Holm} = .026$, $d = 0.73$; SUS: $t = 3.146$, $p_{Holm} = .012$, $d = 0.81$;) und ILIAS und OPAL (UEQ: $t = 2.306$, $p_{Holm} = .057$, n.s., $d = 0.60$; SUS: $t = 2.677$, $p_{Holm} = .025$, $d = 0.69$) hin. Excel und ILIAS unterscheiden sich in der pragmatischen Qualität hingegen nicht (UEQ: $t = 0.521$, $p_{Holm} = .607$, $d = 0.13$; SUS: $t = 0.469$, $p_{Holm} = .643$, $d = 0.12$).

Tabelle 3: Pearson-Korrelationskoeffizienten für Excel

		1.	2.	3.	4.	5.
1. Erfolgsrate	*r* *p*	— —				
2. Bearbeitungszeit	*r* *p*	-.723 .002	— —			
3. SUS	*r* *p*	.540 .038	-.321 .243	— —		
4. ATT (UEQ)	*r* *p*	.204 .467	-.251 .367	.651 .009	— —	
5. PQ (UEQ)	*r* *p*	.546 .035	-.307 .266	.835 < .001	.728 .002	— —
6. HQ (UEQ)	*r* *p*	.124 .661	.162 .563	.432 .108	.449 .093	.582 .023

5.2 Erfolgsrate und Bearbeitungszeit

Über alle drei Aufgaben gemittelt weist der Friedman-Test zunächst auf Unterschiede in der Erfolgsrate zwischen den getesteten Systemen hin ($\chi^2 = 6.16$, $p = .046$). Alle Post-hoc-Tests überschreiten jedoch nach Korrektur für multiples Testen das notwendige Signifikanzniveau. Die Bearbeitungszeit zwischen den Systemen differiert deutlicher ($F(2, 28) = 5.814$, $p = .008$). Die gewählten Testaufgaben konnten mit Excel signifikant schneller bearbeitet werden als mit ILIAS ($t = -3.074$, $p_{Holm} = .014$, $d = -0{,}79$) und OPAL ($t = -2.815$, $p_{Holm} = .018$, $d = -0{,}72$). Während die Testpersonen für die Bearbeitung mit Excel im Mittel etwa 6 Minuten brauchten, dauerte die Bearbeitung in ILIAS und OPAL fast 2 Minuten länger (siehe Abb. 1). Bei ILIAS und OPAL ist die mittlere Bearbeitungszeit hingegen nahezu identisch ($t = 0.260$, $p_{Holm} = .797$, $d = -0.07$).

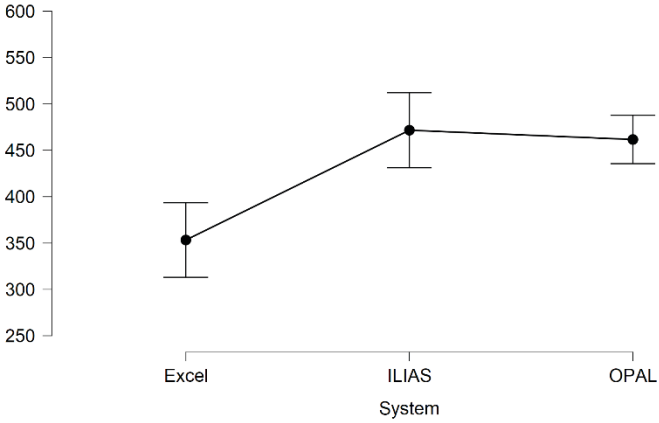

Abb. 1: Mittlere Bearbeitungszeiten in Sekunden über alle Testaufgaben summiert.

6 Diskussion

Die Ergebnisse des Nutzertests geben erste Hinweise darauf, dass sich typische Aufgaben beim Itempool-Management mit einem tabellarischen User Interface, wie es Excel anbietet, effektiver und effizienter bearbeiten lassen, als mit den für LMS typischen formularähnlichen Benutzerschnittstellen. Den Testpersonen gelang nicht nur die Bearbeitung der ausgewählten Aufgaben deutlich schneller. Auch subjektiv erlebten die Testpersonen das Arbeiten mit Excel gegenüber den beiden getesteten Vergleichssystemen als zufriedenstellender. Entwickler von LMS könnten daher einen Wechsel oder eine ergänzende tabellarische Itembearbeitung innerhalb ihrer Systeme in Betracht ziehen. Mit einem mittleren SUS-Score von 65.8 erreicht Excel in der vorliegenden Studie einen noch akzeptablen Wert (Bangor et al., 2008), bleibt aber weit unter dem Niveau populärer Onlinedienste (Kortum & Bangor, 2013). Auch die gegenüber den Vergleichssystemen nicht wesentlich besseren Erfolgsraten weisen auf noch bestehende Probleme hin. Weitere Entwicklungsschritte erscheinen daher notwendig. Beispielsweise können bei vielen Antwortoptionen Tabellen sehr breit werden. Das macht dann ein horizontales Scrollen notwendig, was im Sinne einer gebrauchstauglichen Gestaltung zu vermeiden ist (Nielsen, 2005). In Vorbereitung befindet sich deshalb eine gestapelte Ansicht, in der Antwortoptionen eines Items mit den dazugehörigen Attributen (Punkte, negative Punkte) zeilenweise ausgegeben werden.

Als Einschränkung der vorliegenden Studie ist zu benennen, dass die Testpersonen über deutlich mehr Erfahrung im Umgang mit Excel als mit den beiden Vergleichssystemen verfügten. Das könnte sich auf die Bearbeitung der Testaufgaben ausgewirkt haben. Bekannt ist, dass Vorerfahrung mit einem System typischerweise zu einer besseren Bewertung führt (Sauro, 2011). Die Validität der Untersuchung gefährdet das jedoch nicht, da die Orientierung an Fähigkeiten und Erfahrungen eben gerade zu den Kernmerkmalen eines nutzerorientierten Designprozesses gehört. Ist also bekannt, dass Lehrpersonen an Hochschulen über ein vergleichsweise hohes Maß an Erfahrung mit Excel verfügen, so erscheint es sinnvoll, typische Interaktionsschemata auch auf Lern- und Prüfungssysteme zu übertragen.

Insgesamt empfiehlt sich das neue Interaktionskonzept vor allem für häufige und umfangreichere Arbeiten in großen Itempools und damit als Ergänzung zu – und nicht als Ersatz für – bestehende Interaktionskonzepte.

Literatur

Alberola, C., Brau, H., & Walter, G. (2017). Die Kürzung des User Experience Questionnaire UEQ. Mensch und Computer 2017 – Tagungsband, 37–48. https://doi.org/10/gf4zr8

Bangor, A., Kortum, P., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. International Journal of Human-Computer Interaction, 24(6), 574–594. https://doi.org/10/b344sc

Brooke, J. (1996). SUS – a „quick and dirty“ usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Hrsg.), Usability Evaluation in Industry (S. 189–194). London: Taylor & Francis.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2. Aufl.). Hillsdale, NJ: Erlbaum.

Diefenbach, S., & Hassenzahl, M. (2017). Psychologie in der nutzerzentrierten Produktgestaltung. Berlin: Springer.

DIN EN ISO 9241-11. (1998). Ergonommische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten – Teil 11: Anforderungen an die Gebrauchstauglichkeit – Leitsätze (ISO 9241-11:1998); Deutsche Fassung EN ISO 9241-11:1998. Berlin: Beuth.

DIN EN ISO 9241-210. (2010). Ergonomie der Mensch-System-Interaktion – Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme (ISO 9241-210:2010); Deutsche Fassung EN ISO 9241-210:2010. Berlin: Beuth.

Gould, J. D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. Communications of the ACM, 28(3), 300–311. https://doi.org/10/fqgjgq

Hassenzahl, M. (2007). The hedonic/pragmatic model of user experience. In E. L.-C. Law, A. Vermeeren, M. Hassenzahl, & M. Blythe (Hrsg.), Towards a UX manifesto (S. 10–14). Lancaster: COST.

Kortum, P., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. International Journal of Human-Computer Interaction, 29(2), 67–76. https://doi.org/10/gf4zxd

Michel, L. P., Goertz, L., Radomski, S., Fritsch, T., & Baschour, L. (2015). Digitales Prüfen und Bewerten im Hochschulbereich. Berlin: Hochschulforum Digitalisierung.

Nielsen, J. (2005). Scrolling and Scrollbars. Nielsen Norman Group: https://www.nngroup.com/articles/scrolling-and-scrollbars/

Nielsen, J. (2012). How many test users in a usability study? Nielsen Norman Group: https://www.nngroup.com/articles/how-many-test-users/

Pengel, N., Hawlitschek, P., & Karapanos, M. (2019). Ökonomie und Fairness von Constructed-Response-Items in E-Assessments. In T. Köhler, E. Schoop, & N. Kahnwald (Hrsg.), Gemeinschaften in neuen Medien. Erforschung der digitalen Transformation in Wissenschaft, Wirtschaft, Bildung und öffentlicher Verwaltung (S. 101–111). Dresden: TUDpress.

Rauer, M. (2011). Quantitative Usability-Analysen mit der System Usability Scale (SUS). Seibert Media Weblog: https://blog.seibert-media.net/blog/2011/04/11/usablility-analysen-system-usability-scale-sus/

Sauro, J. (2011). Does prior experience affect perceptions of usability? MeasuringU: http://www.measuringu.com/blog/prior-exposure.php

Sauro, J., & Lewis, J. R. (2016). Quantifying the user experience: practical statistics for user research (2. Aufl.). Cambridge, MA: Morgan Kaufmann.

Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. Journal of Personality and Social Psychology, 80(2), 325–339.

Thor, A., Pengel, N., & Wollersheim, H.-W. (2017). Digitalisierte Hochschuldidaktik: Qualitätssicherung von Prüfungen mit dem E-Assessment-Literacy-Tool EAs.LiT. In C. Igel, C. Ullrich, & W. Martin (Hrsg.), Bildungsräume 2017 (S. 179–184). Bonn: Gesellschaft für Informatik.