Aufwandsanalyse für computerunterstützte Multiple-Choice Papierklausuren

Zusammenfassung

Einleitung: Multiple-Choice-Klausuren spielen immer noch eine herausragende Rolle für fakultätsinterne medizinische Prüfungen. Neben inhaltlichen Arbeiten stellt sich die Frage, wie die technische Abwicklung optimiert werden kann. Für Dozenten in der Medizin gibt es zunehmend drei Optionen zur Durchführung von MC-Klausuren: Papierklausuren mit oder ohne Computerunterstützung oder vollständig elektronische Klausuren. Kritische Faktoren sind der Aufwand für die Formatierung der Klausur, der logistische Aufwand bei der Klausurdurchführung, die Qualität, Schnelligkeit und der Aufwand der Klausurkorrektur, die Bereitstellung der Dokumente für die Einsichtnahme, und die statistische Analyse der Klausurergebnisse.

Methoden: An der Universität Würzburg wird seit drei Semestern ein Computerprogramm zur Eingabe und Formatierung der MC-Fragen in medizinischen und anderen Papierklausuren verwendet und optimiert, mit dem im Wintersemester (WS) 2009/2010 elf, im Sommersemester (SS) 2010 zwölf und im WS 2010/11 dreizehn medizinische Klausuren erstellt und anschließend die eingescannten Antwortblätter automatisch ausgewertet wurden. In den letzten beiden Semestern wurden die Aufwände protokolliert.

Ergebnisse: Der Aufwand der Formatierung und der Auswertung einschließlich nachträglicher Anpassung der Auswertung einer Durchschnittsklausur mit ca. 140 Teilnehmern und ca. 35 Fragen ist von 5-7 Stunden für Klausuren ohne Komplikation im WS 2009/2010 über ca. 2 Stunden im SS 2010 auf ca. 1,5 Stunden im WS 2010/11 gefallen. Einschließlich der Klausuren mit Komplikationen bei der Auswertung betrug die durchschnittliche Zeit im SS 2010 ca. 3 Stunden und im WS 10/11 ca. 2,67 Stunden pro Klausur.

Diskussion: Für konventionelle Multiple-Choice-Klausuren bietet die computergestützte Formatierung und Auswertung von Papierklausuren einen beträchtlichen Zeitvorteil für die Dozenten im Vergleich zur manuellen Korrektur von Papierklausuren und benötigt im Vergleich zu rein elektronischen Klausuren eine deutlich einfachere technische Infrastruktur und weniger Personal bei der Klausurdurchführung.

Schlüsselwörter: Multiple-Choice Prüfungen, Automatisierte Prüfungskorrektur, Aufwandsanalyse

Einleitung

Multiple-Choice (MC) Klausuren spielen immer noch eine herausragende Rolle für medizinische Prüfungen [8]. Neben inhaltlichen Arbeiten [9], [2] stellt sich die Frage, wie die technische Abwicklung optimiert werden kann. Es gibt drei grundlegende Optionen zur Durchführung von MC-Klausuren: Papierklausuren mit oder ohne Computerunterstützung oder elektronische Klausuren:

- A. Traditionell erstellt der Dozent eine Papierklausur mit einem Textverarbeitungssystem, druckt die Klausurbögen, korrigiert die Antworten per Hand und überträgt die Ergebnisse in ein Tabellenkalkulationsprogramm, das die Noten berechnet.
- B. Eine verbesserte Option, für die es bereits kommerzielle Software günstig zu kaufen gibt, benutzt Computer zum Einscannen der Antworten und zur automatischen Zusammenstellung der Ergebnisse im Tabellenkalkulationsprogramm.
- C. Eine weitergehende Automatisierung ist möglich, wenn die Studierenden die Klausuren direkt am Computer schreiben, die Ergebnisse auf einen Server übertragen und automatisch ausgewertet werden.

Die Entscheidung für die ökonomischste Alternative hängt sowohl von der technischen Ausstattung als auch vom gewählten Prozessmodell ab, wobei auch das Risiko

Alexander Mandel¹ Alexander Hörnlein² Marianus Ifland² Edeltraud Lüneburg² Jürgen Deckert¹ Frank Puppe²

1 Universität Würzburg, Medizinische Fakultät, Studiendekanat, Würzburg, Deutschland
2 Universität Würzburg, Fakultät für Mathematik und Informatik, Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik, Würzburg, Deutschland
technischen Versagens zu berücksichtigen ist. In diesem Beitrag untersuchen wir die Effizienz der Durchführung von papierbasierten Klausuren mit Computerunterstützung (B) und vergleichen diese mit den anderen beiden Alternativen A und C. Im Gegensatz zu B gibt es zu C zahlreiche Publikationen (z.B. [6], [3]), darunter auch Aufwandsanalysen, wobei teilweise verschiedene Hardwarevarianten elektronischer Prüfungsorganisationen verglichen werden (Nutzung von Laptops der Studierenden vs. Nutzung hochschuleigener Computer in einem speziellen Testcenter bzw. verteilte in CIP-Pools vs. komplettes Outsourcing an eine Firma).

In einigen Publikationen werden auch Referenzwerte für Aufwände und Kosten konventioneller Prüfungen zum Vergleich angegeben, auf die wir im Folgenden eingehen. In [7] werden die Gesamtkosten, die sich aus Investitionskosten, Personalkosten und Druckkosten zusammensetzen, pro schriftlicher Prüfung bei 96 Prüfungen pro Jahr über einen Zeitraum von 3 Jahren für eine einzelne Klausur nach der Option A 1423 €, nach Option B 1072 € und nach Option C 1746 € geschätzt. Bezüglich der Zeitaufwände zeigt Tabelle 1 einen Vergleich der Schätzungen zweier Studien ([7], siehe Tabelle 1 und Tabelle 2 und [1], Tabelle 27).

Tabelle 1: Zusammenfassung und Vergleich der Aufwandsanalyse für die Optionen A, B und C bei der Durchführung von MC-Klausuren in den Studien [7] und [1]. Dabei entsprechen sich grob die Phasen Entwicklungszeit in Studie [7] und die Katalogerstellung in Studie [1] sowie Klausurlayout und Technische Prüfungsvorbereitung in Studie [1] fehlen Angaben zur Option B, dafür ist die Option C in zwei Varianten aufgeteilt: viele Räume mit je 20 Computerarbeitsplätzen bzw. ein Testcenter mit 120 Computerarbeitsplätzen. Bei unterschiedlichen Aufwandschätzungen beider Varianten ist die erste Variante (viele Räume) mit * und die zweite Variante (Testcenter) mit ** gekennzeichnet.

| Studi  | Klausurphasen / Zeit (jeweils in Stunden) | A | B | C |
|--------|------------------------------------------|---|---|---|
| [7]    | Entwicklungszeit 20                      | 21 | 22 |
|        | Katalogerstellung 30                     | 80 |  |
| [1]    | Klausurlayout 3                         | 5  | 1  |
|        | Technische Prüfungsdurchführung 8       | 32°78 ** |
| [7]    | Prüfungsauswertung 2                    | 2  | 0  |
|        | <13,5 ≤ 3,5                             | 0,5 |
| [1]    | Prüfungsauswertung 200                  | 0  |
| Summe  | <38,5 ≤ 31,5                            | 25,5 |
| [1]    | Summe 246                               | 232°7112 ** |

Ein Vergleich beider Schätzungen zeigt ganz erhebliche Diskrepanzen, die wohl teilweise darauf zurückzuführen sind, dass bei [7] von weniger Klausurteilnehmern mit weniger Fragen pro Klausur ausgegangen wird. Insgesamt fällt auf, dass in [1] generell wesentlich höhere Zeitaufwände kalkuliert werden, wobei die Schätzung von 200 Stunden für die manuelle Auswertung bei A (d.h. ½ Minute pro Frage bei 60 Fragen in 400 Klausuren) wohl auf einem Mix von Freitext-Fragen und geschlossenen Fragen basiert, während bei C nur geschlossene Fragen zugrundegelegt werden. Weiterhin fällt auf, dass für die Prüfungsvorbereitung und Prüfungsdurchführung in [1] beträchtliche Aufwände für Funktionsprüfungen der Computer und Fachaufsichten sowie technischem Support zusammenkommen, während in den Schätzungen von [7] diese Faktoren vernachlässigt werden.

Aus beiden Studien kann ein beträchtliches Potential für die Option B abgeleitet werden, wenn es gelingt, die Vor- teile von A mit geringem technischem Aufwand bei der Vorbereitung und Durchführung der Prüfungen und die Vorteile von C einer geringen Korrekturzeit zu kombinieren. Im Folgenden analysieren wir den Zeitaufwand bei computerunterstützten reinen Multiple-Choice-Papierklausuren mit automatischer Korrektur der eingescanneten Antwortblätter. Andere Frageformate, die eine Zahl- und Texteingabe erfordern, können zwar mitverwaltet werden, aber müssten manuell korrigiert werden. Zur Umsetzung wurde an der Universität Würzburg nach Erfahrungen mit dem kostenpflichtigen spidMED-Service des IMPP [https://www.impp.de/spidMED/] (Link geprüft 11.7.2011; Service seit 1.7.11 nicht mehr verfügbar) sowie einem kommerziellen Programm zur Kreuzerkennung von Multi ple-Choice-Klausuren eine Komponente für computerge- stützte Papierprüfungen entwickelt. Da diese Komponente auf einem aus Studiengebühren finanzierten universitätsweiten Framework zur Entwicklung fallbasiert Trainings- systeme aufbaut (vgl. [5], [4]), waren die zusätzlichen Investitionskosten relativ gering. In Abschnitt 2 werden das Prozessmodell und die kritischen Aspekte computer-gestützter Papierprüfungen beschrieben, in Abschnitt 3 die technischen Aufwände für die verschiedenen Phasen der 12 bzw. 13 Klausuren im SS 2010 und im WS 2010/11 präsentiert (ohne Berücksichtigung der inhaltlichen Arbeit) und in Abschnitt 4 die Option B auf qualitativ- tiver Ebene mit den Optionen A und C verglichen.

Methoden und Prozessmodell

Kritische Faktoren bei der Durchführung von Klausuren sind neben der hier nicht im Fokus stehenden inhaltlichen Arbeit der Aufwand für die Formatierung der Klausur, der logistische Aufwand bei der Klausurdurchführung, die Qualität, Schnelligkeit und der Aufwand der Klausurkorrektur, die Bereitstellung der Dokumente für die Einsichtnahme und die statistische Analyse der Klausurergebnisse. Im Folgenden beschreiben wir ein allgemeines Prozessmodell mit verschiedenen Varianten:

Erstellung und Formatierung der Klausur

Die Fragen einer Klausur können von einem oder mehreren Dozenten (z.B. bei Ringvorlesungen) kommen, es können alte Fragesammlungen auf Papier oder aus einer Datenbank wiederverwendet oder die Fragen ganzer oder teilweise neu erstellt werden. Häufig kontrollieren verschiedene Personen die Fragen, so dass es mehrere Iterationen gibt. Die Fragen können sich auf Bilder oder Fallbeschreibungen beziehen, oft gibt es dann auch
Tabelle 2: Informationen und Zeitangaben zum Aufwand bei der Durchführung von 13 bzw. 12 MC-Papierklausuren mit Computerunterstützung im WS 2010/2011 und SS 2010. Im WS 09/10 wurden elf Klausuren durchgeführt, deren Zeitaufwände nicht mit protokolliert wurden. Daher wird in der letzten Zeile nur eine Durchschnittsschätzung für den Zeitaufwand einer typischen Klausur ohne Komplikationen gezeigt. In allen Klausuren muss für das Ausdrucken der Klausuren ca. eine halbe bis eine Stunde sowie für die Klausuraufsicht und das Auslegen der Klausuren ca. gut zwei Stunden (2 Aufsichtspersonen bei ca. einer Stunde Klausurzeit) hinzugezählt werden.

mehre zusammenhängende Fragen („key feature Fragen“). Die Antwortalternativen können vom Typ A (Einfachauswahl), Typ X (wahr/falsch) oder PickN (Mehrfachauswahl) sein (vgl. http://www.let.ethz.ch/exam_eval/onlinetests/faq/nomenklatur_fragetypen.pdf). Während bei manueller Klausurkorrektur (A) Dozenten die Fragen meist direkt in einem Textverarbeitungsprogramm formatieren, sind bei B und C indirekte Formate üblich. Entweder können die Fragen aus einer Datenbank selektiert werden oder die Dozenten geben die Fragen in einem bestimmten Format ein und der Computer generiert daraus die Klausurvorlage. Auch dabei gibt es zwei Varianten: entweder die Eingabe über ein Formular oder die Eingabe in einem Textverarbeitungssystem mit Layout-Vorgaben, das erst durch einen Parse-Vorgang in das interne Format überführt wird. Um Abschreiben zu erschweren, werden bei Option A oft zwei bis vier Klausurvarianten durch Vertauschen von Fragen und Antwortalternativen manuell erstellt. Bei Option B und C wird das Vertauschen meist automatisiert, so dass jeder Teilnehmer eine andere Klausurvariante bekommt.

Im Rahmen unserer Studie gab es noch keine Fragebank. Dafür wurden die Dozenten weitgehend von der Formatierung entlastet, indem sie Prüfungen als Word-Datei an einen Koordinator geschickt haben, der die notwendigen Formierungen vorgenommen hat. Es wurden Einfach- und Mehrfachauswahlfragen (Typ A und PickN) verwendet. Letzteres ist in Tabelle 2 durch ein „ja“ in der Spalte „mehrere Antworten pro Frage möglich“ gekennzeichnet. Während es im WS 2009/2010 ein relativ kompliziertes Eingabeformat mit vielen Optionen gab, wurde ab dem darauffolgenden SS 2010 das Eingabeformat an die häufigsten Vorlagen der Dozenten angeglichen und stark vereinfacht. Dieses vereinfachte Format (siehe Abbildung 1) wurde den Dozenten mitgeteilt, um den Koordinator zu entlasten. Allerdings mussten die Dozenten sich nicht daran halten, da der Koordinator nach wie vor die Endredaktion übernommen hat. In unseren Aufwandsmessungen in Abschnitt 3 beginnen wir daher mit einer beliebig formatierten Klausurvorlage und messen als ersten Schritt den Aufwand der Nachformatierung durch den Koordinator.

Prüfungsvorbereitung und Klausurdurchführung

Dazu gehören die Aufwände und Kosten für das Ausdrucken der Klausur sowie die Aufwände für das Auslegen
Prüfungsvorbereitung und Klausur durchführung fasst Aufwände zusammen, die in unserem Modell bei den Dozenten liegen, d.h. das Drucken und Heften mit ca. ½ - 1 Stunde (entweder auf eigenem Drucker mit Heften oder in einem Copy-Shop mit Hin- und Rücktransport) sowie die Vorbereitung und Aufsicht bei der eigentlichen Klausur mit typischerweise zwei Personen für ca. eine Stunde Klausurzeit. Da diese Aufwände von ca. 3 Stunden bei jeder Papierklausur anfallen und unabhängig vom Koordinator sind, werden sie in Tab. 2 nicht gesondert ausgewiesen, aber in der Diskussion berücksichtigt.

Klausurauswertung

Während bei Option A die Dozenten die Klausuren manuell korrigieren und die Daten in ein Tabellenkalkulationsprogramm übertragen und bei Option C der Computer die Rohergebnisse sofort liefert, hängt die Effizienz bei der Option B von der Scan Geschwindigkeit und Qualität ab. Da häufig nachträglich noch einzelne Fragen aus der Wertung genommen werden bzw. das Notenschema angepasst wird, ist in allen Optionen A, B, C die Einfachheit der Anpassung der Auswertung wichtig. Weiterhin werden bei den Optionen B und C verschiedene Statistiken (z.B. Trennschärfe der Fragen) automatisch erzeugt.

Das Hauptaugenmerk dieser Studie dient der genauen Analyse des Zeitaufwandes der Klausurauswertung für die Option B. Dazu wird dieser Schritt in Teilschritte zerlegt:

- Das Scannen umfasst im einfachsten Fall das Einlegen der Antwortbögen in einen Scanner. Bei manchen Klausuren waren die Antwortbögen an die Angabenblätter geheftet oder die Antwortbögen verschiedener Klausuren waren vermischt, so dass sie vorher abgetrennt bzw. sortiert werden mussten. Diese Aufwände

Abbildung 1: WORD-Eingabedokument für eine exemplarische Musterklausur. Das Dokument wird geprüft, in eine interne Datenstruktur überführt und daraus das (randomisierte) Klausurdokument mit Angabenblättern, die sehr ähnlich aussehen, aber natürlich keine fettgedruckte Lösungen enthalten, sowie dem Antwortbogen (s. Abb. 2) generiert.
wurden mitgezählt. Während im WS 09/10 ein Hochleistungsscanner in der Universitätsbibliothek verwendet wurde, der aber wegen der Transportzeiten und der Notwendigkeit von Terminvereinbarungen unpraktisch war, wurde ab dem SS 10 ein preisgünstiger Scanner (ca. 1000 Euro) für die Klausurauswertung beschafft, der allerdings nur über einen Blatteinzug von 50 Blättern verfügte und qualitativ nicht so gut war. Die geringere Scan-Qualität konnte jedoch durch eine bessere Auswertungssoftware (s. nächsten Punkt) kompensiert werden.

- Die Auswertung umfasst die automatische Kreuzerkennung der eingescannten Blätter mit manueller Kontrolle und ggf. Nachbearbeitung. Das Programm zur Kreuzerkennung wurde in jedem der drei betrachteten Semester überarbeitet und jeweils durch eine verbesserte Version ersetzt. Alle Versionen boten eine überblickliche Darstellung zur manuellen Kontrolle an, in der die sicher erkannten Kreuze grün, die als unsicher erkannten Kreuze rot markiert waren und eine gelbe bzw. rosa Markierung verwendet wurde, wenn die Anzahl der erkannten Kreuze größer oder kleiner als die Anzahl der erwarteten Kreuze ist. Die aktuelle Version, die seit dem WS 2010/11 eingesetzt wird, kombiniert drei verschiedene Verfahren zur Kreuzerkennung, was zwar die Laufzeit des Kreuzerkennungsprogramms verlängert, aber den Aufwand der manuellen Nacharbeit deutlich reduziert. Das Ergebnis der Kreuzerkennung ist in allen Versionen eine Excel-Tabelle mit den Bewertungen für jeden Teilnehmer und jede Frage einschl. verschiedener Statistiken wie Trennschärfe sowie Dokumente für die Klausureinsichtnahme. 

- Wenn einzelne Fragen missverständlich formuliert wurden oder aus anderen Gründen angepasst oder aus der Wertung genommen werden müssen, entsteht ein Aufwand zur Anpassung der Auswertung. Obwohl dieser Aufwand inhaltlich bedingt ist, haben wir ihn in Tabelle 2 mit ausgewiesen. 

- Der allgemeine Kommunikationsaufwand wird in Tabelle 2 unter der Spalte „Sonstiges / Support“ aufgeführt. Er nimmt naturgemäß im Laufe der Semester ab, wenn die Dozenten mit dem Vorgehensmodell der Klausurdurchführung vertraut sind, aber ist bei Komplikationen höher.

Bei allen Klausuren in Tabelle 2 mit einer Ausnahme im SS 2010 bekamen die Teilnehmer verschiedene Klausurbögen mit gleichen Fragen, indem die Reihenfolge der Fragen bzw. Antwortalternativen vertauscht wurden („Randomisierung = ja“ in Tabelle 2). Die Wahl dieser Option erfordert Vertrauen in die Technik, da mit randomisierten Klausuren eine manuelle Korrektur sehr aufwändig wäre. Andererseits ist es ein wichtiges Argument für den Einsatz computergestützter Klausuren, da so Abschrei- –ben deutlich erschwert und das Auslegen der Klausurvarianten im Prüfungsraum vereinfacht wird. Eine Übersicht über das Prozessmodell bei der Klausurerstellung und -bearbeitung zeigt Abbildung 3.

**Ergebnisse**

Im WS 09/10 wurden elf, im SS 2010 zwölf und im WS 10/11 dreizehn Multiple-Choice-Papierklausuren in der Medizin mit Computerunterstützung erstellt und ausgewertet. Alle Klausuren bis auf eine waren randomisiert. Während ab dem SS 2010 die Aufwände vom Koordinator protokolliert wurden, gab es für das WS 09/10 vom gleichen Koordinator nur nachträgliche grobe Abschätzungen für eine typische Klausur ohne besondere Komplikationen. Die Ergebnisse zeigt Tabelle 2.

Im SS 2010 und WS 10/11 waren bis auf vier Klausuren alle personalisiert, d.h. für jeden Teilnehmer wurde Name und Matrikelnummer auf der Klausur aufgedruckt (mit Reserve-Klausuren für nicht angemeldete Nachzügler).

Es nahmen im SS 10 bzw. WS 10/11 durchschnittlich 143 bzw. 137 Teilnehmer an einer Klausur teil, die im Schnitt jeweils 37 Fragen umfasste. Knapp die Hälfte der Klausuren erlaubte mehrere Antworten pro Frage, die anderen nur genau eine Antwort. Gemessen wurde der zeitliche Aufwand für den Koordinator, der den Dozenten bei der Klausurerstellung und Auswertung hilft. Der durchschnittliche Zeitaufwand wird entsprechend den Ausführungen in Abschnitt 2 in fünf Bereiche aufgeteilt:

- **Nachbearbeitung der Klausurvorsorge:** Während sie im WS 09/10 noch 2-3 Stunden dauerte, sank die Zeit im SS 10 und WS 10/11 auf nur 49 Minuten; bei komplikationslosen Klausuren sogar auf 32 bzw. 23 Minuten. Hier ist ein weiteres Absinken zu erwarten, da es für die Dozenten nur eine Frage der Gewöhnung ist, welches Format sie an den Koordinator schicken. Je ähnlicher es dem in Abbildung 1 gezeigten (WORD-)EingabefORMAT ist, desto weniger Nacharbeit fällt für den Koordinator an.

- **Scannen:** Der Scan-Aufwand hängt hauptsächlich von der Größe des Blatteinzuges und der Scan-Geschwindigkeit ab. Mit derzeit genutzten relativ einfachen Scannern dauert das Einscannen einer Klausur ohne Komplikationen mit ca. 140 Antwortbögen im günstigsten Fall 20-25 Minuten. Der tatsächliche gemessene Durchschnittswert aller Klausuren lag im WS 10/11...
Abbildung 3: Flussdiagramm für den Workflow zur Klausurerstellung und -bearbeitung. Es zeigt die Dokumente in den Kästchen und die Bearbeitungsschritte bei den Pfeilen. In Klammern sind bei den Dokumenten die Typen (Office-Dokumente Word und Excel, Austauschformate PDF und XML, Papier, Bildverarbeitungsformat TIFF sowie Objekte in der Programmiersprache Java) und bei den Bearbeitungsschritten die Personen (Dozent; hellgrau bzw. orange bzw. Koordinator dunkelgrau bzw. blau hervorgehoben) bzw. das Programm angegeben. Einige Kommunikationsschritte (z.B. das Verschicken von E-Mails) sind nicht angegeben.

bei 28 Minuten und im SS 2010 bei 42 Minuten, was hauptsächlich dadurch bedingt war, dass die Einstellungen am Scanner für jede Klausur angepasst werden mussten, um ein optimales Ergebnis zu erzielen. Die notwendigen Schritte werden mittlerweile durch die Auswertungsoftware erledigt bzw. sind durch den Verzicht auf Grauwerte auf dem Antwortbogen entfallen.

**Auswertung:** Der kritischste Schritt ist die Auswertung der Kreuzerkennung auf den Antwortbögen, da davon die Praktikabilität des ganzen Verfahrens abhängt. Um die Qualität der Kreuzerkennung zu sichern, ist ein manueller Überprüfungsschritt mit Darstellung der erkannten Kreuze in den Ampel-Farben (siehe Abschnitt 2) Teil der Auswertung. Der durchschnittliche Aufwand war im WS 10/11 und im SS 10 jeweils etwa 50 Minuten für Klausuren mit ca. 140 Teilnehmern und 37 Fragen pro Klausur. Da verschiedene Kreuzerkennungssoftware eingesetzt wurde, ist es allerdings aussagekräftiger, den Auswertungsaufwand aller Klausuren zu betrachten, die mit der neuen Kreuzerkennung korrigiert wurden, d.h. alle Klausuren im WS 10/11 außer den beiden Pathologie-Klausuren. Hier hat sich die durchschnittliche Auswertungszeit mit nur 26 Minuten pro Klausur fast halbiert.

**Anpassung der Auswertung:** Der Aufwand hängt von Faktoren ab, die sich durch die Art der Auswertung nicht beeinflussen lassen und geht nur indirekt ein, da die verwendete Software ein Korrigieren des Bewertungsschemas oder ein Herausnehmen einzelner Fragen aus der Wertung relativ einfach machen sollte. Der durchschnittliche Aufwand war im WS 10/11 und im SS 10 jeweils ca. 20 Minuten und lag in den meisten Fällen bei 0. Lediglich in der Klausur Infektiologie im WS 2010/11 war er bedingt durch Besonderheiten der Anpassung mit 180 Minuten ungewöhnlich hoch.

**Sonstiges/Support:** Die allgemeine Kommunikation zusätzlich zu den angegebenen Zeiten lag im WS 10/11 bei 12 Minuten, im SS 10 bei 20 Minuten.

In der Summe ist der Aufwand der Klausurbearbeitung ohne Ausdrucken und Klausuraufsicht mit ca. 140 Teilnehmern und ca. 35 Fragen für den Koordinator von 5-7 Stunden für „gute“ Klausuren ohne Komplikation im WS 2009/2010 über ca. 2 Stunden im SS 2010 auf ca. 1,5 Stunden im WS 2010/11 gefallen. Für die am effizientesten korrigierte Klausur „Allgemeinmedizin“ betrug der Aufwand im WS 10/11 sogar nur 65 Minuten bei 121 Teilnehmern und 30 Fragen. Mit Komplikationen stieg die durchschnittliche Zeit auf 160 bzw. 179 Minuten pro Klausur im WS 10/11 bzw. im SS 2010 im WS 2009/2010 ist die Zahl sehr viel höher und nicht ausgewiesen. Die Zahlen zeigen deutlich, dass die Existenz und die Behandlung von Komplikationen für die durchschnittliche Gesamteffizienz fast genauso wichtig sind wie das Basismodell.

**Diskussion**

Insgesamt lässt sich feststellen, dass im SS 2010 und noch mehr im WS 10/11 die Zeitaufwände für Dozente und Koordinatoren ziemlich gering sind. Obwohl immer Raum für weitere Verbesserungen besteht, dürften die
Durchschnittswerte der 9 komplikationslosen Klausuren im WS 2010/11 schon ziemlich nah am Optimum von ca. 1 bis 1,5 Stunden Zeitaufwand pro Klausur liegen (ohne Berücksichtigung des inhaltlichen Aufwandes). Dazu muss noch die Zeit für das Ausdrucken der Klausur von 0,5 bis 1 Stunde addiert werden. Die Gesamtzeit ist vergleichbar mit dem minimalen Zeitaufwand für die Klausuraufsicht von ca. 2 Stunden, die nicht optimierbar ist. Allerdings wurden diese Zahlen nicht auf Anhieb erreicht, da in der Einführungphase im WS 2009/2010 die Zeitaufwände für Klausuren ohne Komplikationen mit 5 bis 7 Stunden für die Dozenten deutlich größer waren und der Koordinator insbesondere bei Klausuren mit Komplikationen einen beträchtlichen Zusatzaufwand hatte.

Dabei scheinen die Anzahl der Klausurteilnehmer und die Anzahl der Fragen pro Klausur nur einen relativ geringen Einfluss auf den Gesamtaufwand zu haben, da mehr Fragen zwar einen erhöhten Formatierungsaufwand bedingen, und mehr Fragen und mehr Teilnehmer den Scanner- und Auswertungsaufwand erhöhen, der Zusatzaufwand sich im Vergleich zu dem Basisaufwand aber in Grenzen hält. Allerdings lassen die empirischen Daten dazu keine klaren Aussagen zu, da die Klausuren relativ homogen bezüglich Fragen- und Teilnehmerzahl sind und die wenigen Klausuren mit stärkeren Abweichungen Komplikationen hatten und deswegen nicht vergleichbar sind.

Wir greifen aufgrund dieser Daten die Gesamtkostenmodelle für die Optionen A, B und C aus [7] und [1] auf und vergleichen sie auf qualitativer Ebene mit den hier ermittelten Aufwänden der Option B. Im Vergleich der Optionen A und B bleiben die Aufwände für die Erstellung der Klausur und die Klausurdurchführung ungefähr gleich, lediglich bei der Klausurauswertung gibt es Unterschiede: Während die Korrekturzeit bei Option A, die in [7] mit 13,5 Stunden pro Klausur und in [1] noch sehr viel höher geschätzt wurde, in unseren Analysen bei Option B ohne Komplikationen auf ca. 1 bis 1,5 Stunden sinkt, kommen zusätzliche Kosten für den Scanner (ca. 1000 Euro) und für die Entwicklung bzw. Anschaffung der Software und deren Wartung hinzu (die an der Universität Würzburg gering ausfielen, da die Korrektursoftware nur eine zusätzliche Komponente im Rahmen eines großen Blended Learning Projektes ist; s.o.). Im Vergleich der Optionen B und C sind die Korrekturzeiten grob vergleichbar und in beiden Optionen ist die Entwicklung bzw. Anschaffung von Software und deren Wartung notwendig. Während bei B Zeitaufwand für das Erzeugen und Ausdrucken des Klausurdokumentes anfällt (ca. 0,5 Stunden für Klausurformatierung und 0,5 bis 1 Stunde für das Drucken), müssen bei C Tests auf Funktionsprüfungen der Computer berücksichtigt werden, die in [7] nicht ausgewiesen sind, aber in [1] 8 bis 32 Stunden umfassen. Ähnliches gilt für die Prüfungsdurchführung, da bei C zusätzlich zur fachlichen Klausuraufsicht auch technisch versiertes Personal anwesend sein sollte, was bei B entfällt. Weitere Unterschiede gibt es bezüglich der Druckkosten (minimal ca. 56 Euro pro Klausur) und der Scanner-Investition bei Option B im Vergleich zu den Investitionen in die Infrastruktur, die zur Durchführung rein elektronischer Klausuren in Option C notwendig ist. Letztere sind schwerer zu beziffern, da es viele Varianten gibt, die von kompletter Ausstattung eines Prüfungscenters mit eigenen Rechnern bis zu elektronischen Prüfungen auf Laptops der Studierenden reichen. Die Studie [1] deutet in diesem Zusammenhang an, dass geringere Investitionskosten einen (deutlich) höheren Zeitaufwand bedingen. Schließlich muss noch das Risiko und der Schweregrad von Komplikationen betrachtet werden, die bei C deutlich mehr ins Gewicht fallen als bei B.

Daher ist das Ziel, die Dozenten möglichst kostengünstig bei der Korrektur von Multiple-Choice-Klausuren zu entlasten, am besten mit der Option B, d.h. papierbasierten Klausuren mit Computerunterstützung, zu erreichen. Elektronische Klausuren lügen sich derzeit nach unseren Untersuchungen nur dann, wenn auch die Möglichkeiten neuer Aufgabenarten jenseits von konventionellem Multiple-Choice bei der Klausurerstellung genutzt werden, wie z.B. Long-Menu-Fragen oder anderen Fragearten, das Zeigen von Videos, das Arbeiten an virtuellen Mikroskopen oder das interaktive Lösen von Trainingsfällen.

**Interessenkonflikt**

Die Autoren erklären, dass sie keine Interessenskonflikte im Zusammenhang mit diesem Artikel haben.

**Literatur**

1. Bücking J, Schwedes K, Laue H. Computergestützte Klausuren an der Universität Bremen, ZMML (Zentrum für Multimedia in der Lehre), Arbeitsbericht. Bremen: Universität Bremen; 2007. Zugänglich unter/available from: http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf
2. Fischer M, Kopp V. Computer-based pre-clinical assessment: Does the embedding of multiple-choice questions in a clinical context change performance? GMS Z Med Ausbild. 2006;23(3):Doc52. Zugänglich unter/available from: http://www.eimgs.de/static/de/journals/zma/2006-23/zma000271.shtml
3. Frey P. Computerbasiert prüfen: Möglichkeiten und Grenzen. GMS Z Med Ausbild. 2006;23(3):Doc49. Zugänglich unter/available from: http://www.eimgs.de/static/de/journals/zma/2006-23/zma000268.shtml
4. Hörlein A, Ifland M, Klüg P, und Puppe F. Konzeption und Einsatz eines fallbasierten Computerprüfungs instruments in der klinischen Lehre: Akzeptanz der Studierenden. GMS Med Inform Biomed Epidemiol. 2009;5(1):Doc07. DOI: 10.3205/mibep000086
5. Hörlein A, Mandel A, Ifland M, Lüneberg E, Deckert J, Puppe F. Akzeptanz medizinischer Trainingsspiele als Ergänzung zu Vorlesungen. GMS Z Med Ausbild. 2011;28(3):Doc42. DOI: 10.3205/zma000754
6. Kopp V, Herrmann S, Müller T, Vogel P, Liebhardt H, Fischer MR. Einsatz eines fallbasierten Computerprüfungs instruments in der klinischen Lehre: Akzeptanz der Studierenden. GMS Z Med Ausbild. 2005;22(1):Doc11. Zugänglich unter/available from: http://www.eimgs.de/static/de/journals/zma/2005-22/zma000011.shtml
7. Krückeberg J, Paulmann V, Fischer V, Haller H, Matthies H. Elektronische Testverfahren als Bestandteil von Qualitätsmanagement und Dynamisierungsprozessen in der medizinischen Ausbildung. GMS Med Inform Biom Epidemiol. 2008;4(2):Doc08. Zugänglich unter/available from: http://www.egms.de/static/de/journals/mibe/2008-4/mibe000067.shtml

8. Möltner A, Duelli R, Resch F, Schultz JH, Jünger J. Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten. GMS Z Med Ausbild. 2010;27(3):Doc44. DOI: 10.3205/zma000681

9. Smolle J. Klinische MC-Fragen rasch und einfach erstellen – ein Praxisleitfaden für Lehrende. Berlin/New York: Walter de Gruyter; 2008.

Korrespondenzadresse:
Prof. Dr. Frank Puppe
Universität Würzburg, Fakultät für Mathematik und Informatik, Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik, Am Hubland, 97074 Würzburg, Deutschland, Tel.: +49 (0)931/31-86730, Fax: +49 (0)931/31-86732
puppe@informatik.uni-wuerzburg.de

Bitte zitieren als
Mandel A, Hörnlein A, Ifland M, Lüneburg E, Deckert J, Puppe F. Aufwandsanalyse für computerunterstützte Multiple-Choice Papierklausuren. GMS Z Med Ausbild. 2011;28(4):Doc55. DOI: 10.3205/zma000767, URN: urn:nbn:de:0183-zma000767

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2011-28/zma000767.shtml

Eingereicht: 10.09.2010
Überarbeitet: 16.06.2011
Angenommen: 16.06.2011
Veröffentlicht: 15.11.2011

Copyright
©2011 Mandel et al. Dieser Artikel ist ein Open Access-Artikel und steht unter den Creative Commons Lizenzbedingungen (http://creativecommons.org/licenses/by-nc-nd/3.0/deed.de). Er darf vervielfältigt, verbreitet und öffentlich zugänglich gemacht werden, vorausgesetzt dass Autor und Quelle genannt werden.
Cost analysis for computer supported multiple-choice paper examinations

Abstract

Introduction: Multiple-choice-examinations are still fundamental for assessment in medical degree programs. In addition to content related research, the optimization of the technical procedure is an important question. Medical examiners face three options: paper-based examinations with or without computer support or completely electronic examinations. Critical aspects are the effort for formatting, the logistic effort during the actual examination, quality, promptness and effort of the correction, the time for making the documents available for inspection by the students, and the statistical analysis of the examination results.

Methods: Since three semesters a computer program for input and formatting of MC-questions in medical and other paper-based examinations is used and continuously improved at Wuerzburg University. In the winter semester (WS) 2009/10 eleven, in the summer semester (SS) 2010 twelve and in WS 2010/11 thirteen medical examinations were accomplished with the program and automatically evaluated. For the last two semesters the remaining manual workload was recorded.

Results: The cost of the formatting and the subsequent analysis including adjustments of the analysis of an average examination with about 140 participants and about 35 questions was 5-7 hours for exams without complications in the winter semester 2009/2010, about 2 hours in SS 2010 and about 1.5 hours in the winter semester 2010/11. Including exams with complications, the average time was about 3 hours per exam in SS 2010 and 2.67 hours for the WS 10/11.

Discussion: For conventional multiple-choice exams the computer-based formatting and evaluation of paper-based exams offers a significant time reduction for lecturers in comparison with the manual correction of paper-based exams and compared to purely electronically conducted exams it needs a much simpler technological infrastructure and fewer staff during the exam.

Keywords: Educational Measurement (I2.399), Self-Evaluation Programs (I2.399.780), Multiple-Choice Examination, Cost Analysis

Introduction

Multiple Choice (MC) exams still play a prominent role for medical tests [8]. In addition to substantive work [9], [2] the question of how the technical aspects can be optimized appears. There are three basic options for the implementation of MC-exams: exam papers with or without computer support or electronic examinations:

• A. Traditionally, the instructor creates an exam paper with a word processing system that prints out the exam sheets, corrects the answers by hand, and transmits the results to a spreadsheet program that calculates the scores.
• B. A better option, for which there is already commercial software to buy, uses computers to scan the responses and to automatically insert the results in a spreadsheet program.
• C. A further automation is possible if the students directly write their exams on the computer with the results being transmitted to a server and automatically evaluated afterwards.

The decision to use the most economical alternative depends on both the technical equipment as well as the selected process model whereby the risk of technical failure must be considered. In this paper we examine the efficiency of the implementation of paper-based exams with computer support (B) and compare these with the other two alternatives A and C. In contrast to B there are numerous publications about C (e.g. [6], [3]), including cost analysis, whereas some different hardware versions of electronic exam designs are compared (using students’ own laptops vs. using university’s computers in a special testing center or in CIP pools vs. complete outsourcing).
In some publications reference values for expenses and costs are provided for comparison with conventional tests, which we will discuss in the following. In [7], the total costs for a single exam which consist of investment costs, personnel costs and printing costs, are € 1423 with option A, with option B € 1072 and with option C € 1746, under the assumption of 96 examinations per year over a period of 3 years. Concerning the time exposure a comparison of estimates from two studies is shown in Table 1 ([7], Table 1 and Table 2 and [1] Table 27).

Table 1: Summary and comparison of the cost analysis for the options A, B and C for the implementation of MC-exams in studies [7] and [1]. Thereby the phases of “developmental period” in study [7] and “catalog production” in study [1] roughly correspond to each other, as well as the phases “exam layout” and “technical exam preparation”. In study [1] data for option B are missing and option C is divided into two types: many rooms with 20 computers each and a test center with 120 computers. With different effort estimations of both variants the first option (many rooms) is marked with a * and the second option (one test center) with **.

| study | exam phase/exam option | A  | B  | C   |
|-------|-------------------------|----|----|-----|
| [7]   | developmental period    | 20 | 21 | 22  |
| [1]   | Catalogue production    | 30 | 80 |     |
| [7]   | exam layout             | 3  | 5  | 1   |
| [1]   | technical exam preparation | 32 | 120 | 120 |
| [7]   | exam procedure          | 2  | 2  |     |
| [1]   | exam procedure          | 8  | 120 | 120 |
| [7]   | exam evaluation         | 13.5 | 5.5 | 0.5 |
| [1]   | exam evaluation         | 200 |     |     |
| [7]   | sum                     | <38.5 | 31.5 | 25.5 |
| [1]   | sum                     | 248 | 232 | 112 |

A comparison of the two estimates shows significant discrepancies, which are probably partly due to the fact that in [7] less exam participants with fewer questions per exam are considered. Overall it is apparent that in [1] generally much higher time expenses are calculated, whereas the estimation of 200 hours for manual evaluations of A (½ minutes per question with 60 questions in 400 exams) probably bases on a mix of free-text questions and closed questions, while in C only closed questions are used. Furthermore, it is striking that for exam preparation and test execution in [1], considerable efforts for the functionality testing of the computers and technical supervisors as well as technical support occur, while these factors are neglected in the estimates of [7]. Significant potential for option B can be derived from both studies if it is possible to combine the advantages of A (little technical effort in preparing and carrying out the tests) and the advantages of C (minor correction time). In the following, we analyze the time required for computer-based pure multiple-choice paper exams with automatic correction of the scanned answer sheets. Other types of questions that require a number and text input can be co-managed, but they would have to be corrected manually. For implementation, a component for computer-based paper exams had been developed at the University of Wuerzburg after experiences with the paid service of IMPP spidMED [https://www.impp.de/spidMED/ (Link checked 11.7.2011; Service terminated: 1.7.2011)] and a commercial program for optical mark recognition of multiple-choice exams. Since this component is based on a university-wide framework for the development of case-based training systems (see [5], [4]), which is funded by tuition fees, the additional investment costs were relatively low. In Section 2, the process model and the critical aspects of computer-based paper tests are described while in Section 3 the technical effort for the various phases of the 12 resp. 13 exams in SS 2010 and WS 2010/11 is presented (without considering the content of the work) and in Section 4, Option B is compared on a qualitative level with options A and C.

Methods and process model

Critical factors in the implementation of examinations are in addition to the substantial content related work, on which we don’t focus here, the logistics involved in the examination procedure, the quality, speed and cost of exam correction, the provision of documents for inspection and the statistical analysis of exam results. Below, we describe a general process model with several variations:

Creating and formatting of the exam

Questions in a written examination may come from one or more lecturers (e.g. lecture series), old question files on paper or from a database can be reused or the questions can be completely or partially created anew. Often the questions are checked by different persons so that there are several iterations. The questions may relate to images or descriptions of cases and there are often several related questions (“key feature questions”). The answer alternatives can be of type A (single selection), type X (true / false) or PickN (multiple choice) (cf. [http://www.let.ethz.ch/exam_eval/onlinetests/faq/ nomenklatur_fragetypen.pdf]). While in manual exam correction (A) the lecturers usually directly format the questions in a word processing program, indirect formats are common in B and C. Either the questions are selected from a database or the lecturers enter them in a specific format from which the computer generates the exam. There are two variants for this process: either input via a form or input into a word processing system with layout specifications, which is converted by a parse operation in the internal format. To discourage copying among exam participants, there are often two to four variants under option A created by swapping exam questions and answer alternatives manually. Under option B and C, the swapping is mostly automated so that each participant gets a different exam version.

In our study no question database existed yet. The lecturers defined the questions and were largely relieved of the formatting by having exam texts sent as a Word file to a coordinator who made the necessary formatting.
Table 2: Informations and time designations about the effort of the implementation of 13 resp. 12 MC-paper exams with computer support in WS 2010/2011 and SS 2010. In WS 09/10 eleven exams were performed with their efforts not being logged. Therefore, in the last line only the average estimate for the time effort of a typical exam without complications is shown. In all written examinations about a half to one hour for printing the exams and about two hours for examination supervision and the distribution of the exams (two supervisors for about an hour of exam time) must be added.

| Course          | Psychometrics | Neurology | Psychiatry | Urology | Dermatology | Gastroenterology | Infectiology | Interciplinary Oncology | Rheumatology | Pathology for Physicians | Pathology for Biomedics | General Medicine | Obstetrics / Gynaecology | Average |
|----------------|---------------|-----------|------------|---------|-------------|------------------|--------------|-------------------------|--------------|--------------------------|------------------------|-----------------|--------------------------|---------|
| WS 10-11       | yes           | yes       | yes        | yes     | yes         | yes              | yes          | yes                     | yes          | yes                      | yes                    | yes             | yes                      | 137     |
| WS 11-11       | yes           | yes       | yes        | yes     | yes         | yes              | yes          | yes                     | yes          | yes                      | yes                    | yes             | yes                      | 137     |
| Average        | yes           | yes       | yes        | yes     | yes         | yes              | yes          | yes                     | yes          | yes                      | yes                    | yes             | yes                      | 137     |

Figure 1: Word document input for a sample exam. The document is parsed into an internal data structure, from which the (randomized) exam document with information sheets that look very similar, but of course contain no bold solutions, and the answer sheet (see figure 2) are generated.
Single and multiple-choice questions (type A and PickN) were both used. The latter is marked in Table 2 by a "yes" in the column "multiple answers per question." While in the winter semester 2009/2010, a relatively complicated input format with many options was used; from the following summer semester 2010 the input format was aligned with the most common templates of the lecturers and simplified. This simplified format (see Figure 1) was communicated to the lecturers in order to relieve the coordinator. However, the lecturers didn’t have to follow them, as the coordinator was still responsible for the final editing. In our effort measurements in section 3, we therefore begin with an arbitrarily formatted exam text and measure the cost of subsequent formatting by the coordinator as a first step.

Exam preparation and exam management

This includes the expenses and costs for the printing of the exam as well as the efforts for laying out the sheets in the auditorium and the exam supervisor. The printing can be done on own printers or in a copy shop, in the latter case a PDF file is sent and then the printed exams are collected. The exams are usually put on the tables in the auditorium. While an alphabetical seating plan must be created in case of personalized exams so that participants are able to find their exams, in case of non-personalized exams the students write their name and matriculation number on the answer sheets, which has to be transferred into the analysis file afterwards. The exam requires the supervision of one or more persons depending on the number of participants.

At the University of Wuerzburg, the costs for printing typical medical exams can be estimated as follows: At about 140 participants and about 35 questions about 140 * 20 = 2800 pages are printed, which at a cost of 2 cents per copy makes around 56 Euro per exam (which must still be paid by the lecturers; in case of the use of color copies it is accordingly more expensive). In Table 2 the column "personalization" indicates, whether personalized exams were used and the "randomization" column marks, if questions and answers were exchanged automatically in order to impede copying. To simplify the correction, a separate answer sheet (see Figure 2 left as used in summer semester 2010 and right as in winter semester 2010/11) was created on which the numbers of the answer to all questions are marked. After initial experience with scanning in WS 09/10, in SS 2010 much more value was placed on good print quality and the use of pencil and eraser for the prevention of scribbling, which markedly improved the level of automation during the correction (see section exam evaluation). The step exam preparation and implementation summarizes efforts that lie in our model with the lecturer, i.e. printing and stapling with ½ - 1 hour (either on own printers with stapling, or in a copy shop with pick-up time) and the preparation and supervision during the actual exam with typically two people for about an hour exam time. Since these expenses of approximately 3 hours occur for each written exam and are regardless of the coordinator, they are not separately identified in Table 2 but included in the discussion.

Exam evaluation

While in option A the lecturer corrects the exams and manually transfers the data into a spreadsheet program and in option C the computer instantly gives the raw results, the efficiency of option B depends on the scanning speed and quality. Since it’s not uncommon for questions to be subsequently removed from the rating or the rating scheme being adapted, in all the options A, B, C the simplicity of the adaptation of the evaluation is important. Also, at options B and C statistics, e.g. discriminative power ("Trennschärfe"), of the questions are automatically generated.

The main focus of this study is the detailed analysis of the time required for the evaluation of an exam with option B. Therefore this step is broken down into smaller steps:

- Scanning, in the simplest case, consists of the insertion of the answer sheets into a scanner. In some exams the answer sheets were tacked to the information sheets or answer sheets of different examinations were mixed, so that they previously had to be separated and sorted. These efforts were counted. While in WS 09/10 a high performance scanner was used in the university library, which was impractical because of transport times and the need for an appointment, in SS 10 an inexpensive scanner (about 1000 €) was purchased for exam evaluation, which, however, could only hold about 50 sheets at the same time and didn’t possess very good quality. However, the lower scan quality could be compensated by better evaluation software (see next item).
- The analysis includes the automatic recognition of crosses on scanned papers with manual inspection and rework if necessary. The program for optical mark recognition was revised and replaced by an improved version in each of three semesters. All versions offered a clear view for manual checking, in which confidently recognized crosses were marked green, probably recognized crosses were marked with red and yellow or pink markings were used if the number of detected crosses was greater or less than the number of optical marks expected. The current version that is used since WS 2010/11 combines three different methods for optical mark recognition, which, while extending the duration of the optical mark recognition program, clearly reduces the effort of manual rework. In all versions the optical mark recognition results were put in an Excel spreadsheet with the scores for each participant and each question, including various statistics such as discriminative power, as well as documents for the exam inspection.
• If some questions were unclear or needed to be adjusted or taken out of the valuation for other reasons, effort for the adaptation of the assessment scheme occurred. Although this expense is conditional to its content, we have identified it in Table 2.

• The overall communication effort is listed in Table 2 under the column "Other/Support". It will naturally decrease over the semesters, when the lecturers are familiar with the process model of exam implementation, but it is higher if complications arise.

For all exams in Table 2 with one exception in SS 2010, participants were given different exam sheets with the same questions but the order of questions and response alternatives had been interchanged ("randomization = yes" in Table 2). Choosing this option requires trust in the technology, since the manual correction of randomized exams would be very difficult. On the other hand, it is an important argument for the use of computer-based exams, since it obviously impedes copying and simplifies the distribution of the exam in the exam room. An overview of the process model in exam preparation and processing is shown in Figure 3.

Results

In WS 09/10, eleven, in SS 2010 twelve and in WS 10/11 thirteen multiple-choice papers in medicine were created and evaluated with computer assistance. All examinations but one were randomized. While for SS 2010 the coordinator recorded the expenditures, there were only rough estimates for a typical exam without major complications from the same coordinator in WS 09/10. The results are shown in table 2.

In SS 2010 and WS 10/11 all exams but four were personalized, i.e. each participant's name and student number was printed on the exam (with reserve exams for undeclared stragglers). In SS 10 resp. WS 10/11 there was an average of 143 or 137 participants per exam, which included 37 questions at average. Almost half of the exams allowed multiple answers per question, the other ones only one answer. Measured was the time required for the coordinator who helps the lecturers with exam preparation and evaluation. The average time is divided in five areas according to the information in section 2:

• **Postprocessing of the exam template**: while it took 2-3 hours in WS 09/10, the time decreased in SS 10 and WS 10/11 to only 49 minutes, in uncomplicated exams even to 32 or 23 minutes. Here, a further drop is expected because it is just a matter of time for the lecturers to get used to the format they send to the coordinator. The more similar it is to the one shown in figure 1 (WORD input format), the less rework is due for the coordinator.

• **Scanning**: The scanning effort mainly depends on the size of the sheet feeder and scanning speed. With the currently used, relatively simple scanner, scanning an exam with about 140 answer sheets without complications takes 20-25 minutes at best. The average of all exams was actually measured at 28 minutes in WS 10/11 and 42 minutes in SS 2010, which was mainly due to the fact that the scanner settings had to be adapted for each exam in order to achieve an optimal result. The necessary steps are now either carried out by the analysis software or are unnecessary because the answer sheet contains no gray values.

• **Evaluation**: The most critical step is the evaluation of the optical mark recognition on the answer sheet because this determines the practicality of the whole process. To ensure the quality of the optical mark recognition, a manual verification step with representation of the detected crosses in the traffic-light colors (see section 2) is part of the evaluation. In SS 10 and WS 10/11 the average expenditure was about 50 minutes per exam with approximately 140 participants and 37 questions each. Since different optical mark recognition programs were used, it is more informative to consider the evaluation effort of all examinations which were corrected with the new optical mark recognition, i.e. all exams in WS 10/11 excluding the two pathology exams. Here, the average evaluation time has almost halved with only 26 minutes per exam.

• **Adaptation of the assessment scheme**: The costs depend on factors that can’t be influenced by the method of evaluation and are addressed only indirectly, because the software used should make the correction of the assessment scheme or the removal of individual questions from the valuation relatively simple. In SS 10 and WS 10/11 the average expenditure was approximately 20 minutes and in most cases at 0. Only in the exam Infectious Diseases in WS 2010/11 it was unusually high with 180 minutes due to special circumstances.

• **Miscellaneous/Support**: The general communication in addition to the indicated times was 12 minutes in WS 10/11 and 20 minutes in SS 10.

In sum, the effort of the exam process without printing and exam supervision for an exam with about 140 participants and about 35 questions has decreased for the
It seems that the number of exam participants and the number of questions per exam have a relatively small influence on the overhead costs because although more questions require increased formatting effort, and more questions and more participants increase the scanning and evaluation effort, the additional effort in comparison to the basic effort is limited. However, the empirical data don’t allow clear statements because the exams are relatively homogeneous concerning the number of questions and participants and the exams with deviations had complications and therefore were not comparable.

Based on this data, we pick up the cost models for the options A, B and C from [7] and [1] and compare them on a qualitative level with our outlined expenses of option B. In the comparison of options A and B the efforts for the creation of the exam and for the exam procedure remain about the same. The differences result from the exam evaluation: The correction time for option A, which was estimated 13.5 hours per exam in [7], and in [1] even longer, drops to 1 to 1.5 hours in our analysis of option B when there are no complications. Additional costs for the scanner (about 1000 €) and the development or acquisition of software and its maintenance (which were very low at the University of Wuerzburg, because the software used for correction was only one additional component of a large blended learning project, see above) have to be added. In comparison of options B and C, the correction times are roughly comparable and in both options the development or acquisition of software and its maintenance is necessary. While in B time expenses for the creation and printing of the document examination incur (about 0.5 hours for the formatting of the exam and 0.5 to 1 hour for printing), at C
function tests of the computers must be included, which aren’t reported in [7] but include 8 to 32 hours in [1]. The same applies to the exam procedure, as with C additionally to the exam supervision technically skilled personnel should be present, which does not apply to B. There are other differences regarding the printing costs (a minimum of 56 € per exam) and scanner investment in option B in comparison with investments in infrastructure, which is necessary for the realization of purely electronic exams in Option C. The latter are difficult to quantify, since there are many variations, ranging from a fully equipped test center with own computers to the use of student’s laptops. In this context the study [1] suggests that lower investment costs lead to (much) more time effort. After all, the risk and severity of complications, which carry much more weight in C than in B, also have to be considered. Therefore, the aim to relieve the lecturers as inexpensively as possible in the correction of multiple-choice exams can be achieved in the best way with option B, which means paper-based exams with computer support. According to our investigations, electronic exams currently only pay off when the opportunities of new types of tasks beyond conventional multiple-choice will be used, such as long-menu questions or other types of questions, the showing of videos, working with virtual microscopes or solving interactive training cases.

Competing interests

The authors declare that they have no competing interests.

References

1. Bücking J, Schwedes K, Laue H. Computergestützte Klausuren an der Universität Bremen, ZMML (Zentrum für Multimedia in der Lehre), Arbeitsbericht, Bremen: Universität Bremen; 2007. Zugänglich unter/available from: http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf
2. Fischer M, Kopp V. Computer-based pre-clinical assessment: Does the embedding of multiple-choice questions in a clinical context change performance? GMS Z Med Ausbild. 2006;23(3):Doc52. Zugänglich unter/available from: http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf
3. Frey P. Computerbasiert prüfen: Möglichkeiten und Grenzen. GMS Z Med Ausbild. 2006;23(3):Doc49. Zugänglich unter/available from: http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf
4. Hörnlein A, Ifland M, Klügi P, und Puppe F. Konzeption und Evaluation eines failbasierten Trainingssystems im universitätsweiten Einsatz (CaseTrain). GMS Med Inform Biom Epidemiol. 2009;5(1):Doc07. DOI: 10.3205/mibe000268
5. Hörnlein A, Mandel A, Ifland M, Lüneberg E, Deckert J, Puppe F. Akzeptanz medizinischer Trainingsfälle als Ergänzung zu Vorlesungen. GMS Z Med Ausbild. 2011;28(3):Doc42. DOI: 10.3205/zma000754
6. Kopp V, Herrmann S, Müller T, Vogel P, Liebhardt H, Fischer MR. Einsatz eines failbasierten Computerprüfungsinstruments in der klinischen Lehre: Akzeptanz der Studierenden. GMS Z Med Ausbild. 2005;22(1):Doc11. Zugänglich unter/available from: http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf
7. Krückeberg J, Paulmann V, Fischer V, Haller H, Matthies H. Elektronische Testverfahren als Bestandteil von Qualitätsmanagement und Dynamisierungsprozessen in der medizinischen Ausbildung. GMS Med Inform Biom Epidemiol. 2008;4(2):Doc08. Zugänglich unter/available from: http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf
8. Möltner A, Duelli R, Resch F, Schultz JH, Jünger J. Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten. GMS Z Med Ausbild. 2010;27(3):Doc44. DOI: 10.3205/zma000681
9. Smolle J. Klinische MC-Fragen rasch und einfach erstellen – ein Praxisleitfaden für Lehrende. Berlin/New York: Walter de Gruyter; 2008.

Corresponding author:
Prof. Dr. Frank Puppe
Wuerzburg University, Faculty of Mathematics and Computer Science, Chair of Artificial Intelligence and Applied Informatics, Am Hubland, 97074 Wuerzburg, Germany, Phone: +49 (0)931/31-86730, Fax: +49 (0)931/31-86732
puppe@informatik.uni-wuerzburg.de

Please cite as
Mandel A, Hörnlein A, Ifland M, Lüneburg E, Deckert J, Puppe F. Cost analysis for computer supported multiple-choice paper examinations. GMS Z Med Ausbild. 2011;28(4):Doc55. DOI: 10.3205/zma000767, URN: urn:nbn:de:0183-zma0007672

This article is freely available from http://www.eassessment.uni-bremen.de/documents/eKlausurenBerichtZMML.pdf

Received: 2010-09-10
Revised: 2011-06-16
Accepted: 2011-06-16
Published: 2011-11-15

Copyright ©2011 Mandel et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en). You are free: to Share — to copy, distribute and transmit the work, provided the original author and source are credited.