DISSERTATION

Wiederverwendung und Veröffentlichung von Multiple-Choice-Fragen in medizinischen Prüfungen – Auswirkungen auf psychometrische Kennwerte und Klausurergebnisse

Reuse and Disclosure of Multiple-Choice Questions in Medical School Exams – Effects on Item Psychometrics and Test Results

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Stefan Appelhaus

Erstbetreuung: Prof. Dr. Liane Schenk

Datum der Promotion: 29.11.2024

Inhaltsverzeichnis

Tabellenverzeichnis iii

Abbildungsverzeichnis iv

Abkürzungsverzeichnis v

Zusammenfassung 1

1 Einleitung 4

1.1 Kompetenzbasierte Curricula und Test-Enhanced Learning 4

1.2 Verwendung von MC-Fragen 5

1.2.1 Psychometrische Parameter 5

1.2.2 Vor- und Nachteile von MC-Fragen 6

1.3 Veröffentlichung und Wiederverwendung von Prüfungsitems 7

1.4 Forschungsstand 8

1.5 Ausgangsposition und Fragestellung dieser Studie 8

2 Methodik 11

2.1 Rahmenbedingungen 11

2.1.1 Gleitklausel 11

2.2 Untersuchte Parameter 12

2.3 Analyse 13

3. Ergebnisse 15

3.1 Unterschiede zwischen den Itemgruppen 15

3.2 Veränderung zur Erstverwendung 16

3.3 Veränderung der Itemparameter und Noten über die Zeit 17

4. Diskussion 20

4.1 Kurze Zusammenfassung der Ergebnisse 20

4.2 Interpretation der Ergebnisse 20

4.3 Einbettung der Ergebnisse in den bisherigen Forschungsstand 21

4.4 Stärken und Schwächen der Studie 22

4.5 Implikationen für Praxis und zukünftige Forschung 23
5. Schlussfolgerungen 25
Literaturverzeichnis 26
Eidesstattliche Versicherung 30
Anteilserklärung an den erfolgten Publikationen 31
Druckexemplar der Publikation 32
Lebenslauf 39
Komplette Publikationsliste 42
Danksagung 43

Tabellenverzeichnis

Tabelle 1: Detaillierte Übersicht über die Prüfungsnoten, Bestehensquoten, Verwendung der Gleitklausel, Schwierigkeits- und Trennschärfekoeffizienten im Studienzeitraum. Quellenangabe: Zeilen 5 und 6 aus Appelhaus et al., 2023 (10): 19

Abbildungsverzeichnis

Abbildung 1: Übersicht über eingeschlossene Items. Quellenangabe: Appelhaus et al., 2023 (10) 15

Abbildung 2: Schwierigkeitskoeffizienten (A) und Trennschärfe (B) der einzelnen Itemgruppen. Quellenangabe: Appelhaus et al., 2023 (10) 16

Abbildung 3: Veränderung der Schwierigkeitskoeffizienten (A) und Trennschärfe (B) wiederverwendeter, nicht veröffentlichter und veröffentlichter Items im Vergleich zur ersten Verwendung. Quellenangabe: eigene Abbildung. 17

Abbildung 4: Veränderung des Anteils der wiederverwendeten, veröffentlichten Items (A) der mittleren Note bestandener Studierender (B) der Bestehensquote und des Anteils der von der Gleitklausel betroffenen Studierenden (C) sowie der mittleren Schwierigkeit und Trennschärfe im Studienzeitraum (D). Quellenangabe: Diagramme A und D aus Appelhaus et al., 2023 (10).....18

Abkürzungsverzeichnis

ÄApprO – Approbiationsordnung für Ärzte

ANOVA – Analysis of Variance (Varianzanalyse)

CBME – Competency-based Medical Education

IMPP – Institut für medizinische und pharmazeutische Prüfungsfragen

MC – Multiple Choice

MCQ – Multiple Choice Question

OSCE – Objective Structured Clinical Examination

SoSe - Sommersemester

USA – United States of America

USMLE – United States Medical Licensure Examination

WiSe - Wintersemester

Zusammenfassung

Hintergrund: Formative Prüfungen in kompetenzbasierten Curricula verlangen nach detailliertem Feedback für die Lernenden, was eine zumindest teilweise Veröffentlichung der Prüfungsitems erfordert. Zudem könnte eine Veröffentlichung Prüfungsangst lindern und Transparenz erhöhen. Viele Fakultäten sind aus Ressourcengründen jedoch auf die Wiederverwendung von Prüfungsitems angewiesen, sodass eine Veröffentlichung der Items die Reliabilität, Schwierigkeit und Vergleichbarkeit der nachfolgenden Prüfungen gefährden könnte. Nach sorgfältiger Abwägung entschied sich die Charité-Universitätsmedizin Berlin ab dem Jahr 2017 mittels einer Änderung der Geheimhaltungsmaßnahmen zur de facto-Veröffentlichung der verwendeten Items.

Zielsetzung: Prüfungsverantwortliche stehen vor der Herausforderung, einen Mittelweg zwischen Feedback, Transparenz und Reliabilität zu finden. Unser Szenario bot die bisher einmalige Gelegenheit, den Einfluss von Veröffentlichung und Wiederverwendung von Items auf psychometrische Parameter und Prüfungsergebnisse anhand einer großen Kohorte zu untersuchen und Daten zur Entscheidungsunterstützung zu liefern.

Methodik: Wir untersuchten 5 Klausurperioden 2017 bis 2019 retrospektiv und verglichen die drei Gruppen „Neue Items“, „wiederverwendete, nicht veröffentlichte Items“ und „wiederverwendete, veröffentliche Items“ bezüglich Schwierigkeit und Trennschärfe sowie die Differenz dieser Parameter bei wiederverwendeten Items zu deren Erstverwendung. Wir analysierten die Veränderung von durchschnittlicher Schwierigkeit und Trennschärfe, Prüfungsnoten, Bestehensquoten und Anteil der von der Gleitklausel betroffenen Studierenden über den Studienzeitraum. Zur Analyse dienten einfaktorielle ANOVAs und t-Tests.

Ergebnisse: Wir analysierten 10.148 Items aus 199 Prüfungen mit 23.507 Teilnehmenden. Wiederverwendete, veröffentlichte Items waren leichter *(M = 0,83)* als wiederverwendete, nicht veröffentlichte *(M = 0,71)* und neue Items *(M = 0,66; p* < 0,001*)*. Während des Studienzeitraums stieg der Anteil der wiederverwendeten, veröffentlichten Items kontinuierlich bis auf 48%, der Schwierigkeitskoeffizient veränderte sich von *M = 0,70* auf maximal *0,76* (*p* < 0,001) und die Durchschnittsnote von *M = 2,64* auf *2,41* (*p* < 0,001). D.h., Items wurden leichter und Durchschnittsnoten besser. Die Trennschärfe verbesserte sich minimal, die Bestehensquote war nicht beeinflusst. Der Anteil der von der Gleitklausel betroffenen Studierenden sank von 39,7 auf 26,9 % (*p* < 0,001).

Schlussfolgerung: Die Ergebnisse zeigen, dass wiederverwendete, veröffentlichte Items leichter zu lösen sind und damit die Reliabilität der gesamten Prüfung negativ beeinflussen könnten. Bei Verwendung einer großen Anzahl an Items, wie in unserem Fall, sowie Anwendung der Gleitklausel scheint der Effekt auf die letztendlichen Noten und die Bestehensquoten jedoch gering zu sein, sodass die Vorteile einer Veröffentlichung von Items je nach Zweck der Prüfung die Nachteile überwiegen könnten.

Abstract

Background: Formative assessment in competency-based curricula requires detailed feedback for learners, which requires at least partial disclosure of the examination items. Disclosure could also reduce student anxiety and improve transparency. However, many medical schools rely on the reuse of items for resource reasons, so that disclosure could endanger reliability, difficulty and comparability of future exams. From 2017, after careful consideration, Charité-Universitätsmedizin Berlin decided to de facto publish the items used by changing security measures.

Objective: Examiners are faced with the challenge of finding a balance between feedback, transparency and reliability. Our scenario provided a unique opportunity to analyse the impact of disclosure and reuse on item psychometrics and exam results using a large cohort and to provide decision support data.

Methods: We analysed 5 exam periods from 2017 to 2019 retrospectively and compared the three groups "new items", "reused, not disclosed items" and "reused, disclosed items" in terms of difficulty and discriminatory power as well as the difference between these parameters for reused items and their initial use. We analysed the change in mean difficulty and selectivity, exam grades, pass rates and the proportion of students affected by the sliding clause over the study period. One-way ANOVAs and t-tests were used for the analysis.

Results: We analysed 10,148 items used in 199 examinations with 23,507 participants. Reused, published items were easier (M = 0.83) than reused, unpublished (M = 0.71) and new items (M = 0.66; $p < 0.001$). During the study period, the proportion of reused, published items increased continuously to 48%, the difficulty coefficient changed from M = 0.70 to a maximum of 0.76 ($p < 0.001$) and the mean exam grade from M = 2.64 to 2.41

($p < 0.001$). I.e., items became easier and grades improved. Discrimination improved only slightly, pass rate was not affected. The proportion of students affected by the automatic adjustment clause clause decreased from 39.7 to 26.9 % ($p < 0.001$).

Conclusion: The results show that reused, published items are easier to solve and could therefore negatively influence the reliability of the entire exam. However, when using a large item-bank, as in our study, and applying the automatic adjustment clause, the effect on final exam grades and pass rates appears to be small to non-existent, so that the advantages of disclosing items may outweigh the disadvantages, depending on the purpose of the exam.

1 Einleitung

1.1 Kompetenzbasierte Curricula und Test-Enhanced Learning

Moderne, kompetenzbasierte medizinische Studiengänge (Competency-based Medical Education, CBME) haben zum übergeordneten Ziel, Studierenden die notwendigen Kompetenzen zu vermitteln, um eine sichere, effektive und patientenzentrierte Versorgung zu gewährleisten (1–3). Teil dieser Fokusverschiebung ist eine Veränderung der Prüfungspraxis von einer abprüfenden, Noten vergebenden, *summativen* Prüfung hin zu einer *formativen* Prüfung, die zwar einen gewissen Mindeststandard festlegt, aber insbesondere dazu dienen soll, den oder die Geprüfte zum Lernen anzuregen und anschließend detailliert über individuelle Stärken, Schwächen und Verbesserungspotential zu informieren. In der Literatur findet sich auch oft der Vergleich zwischen dem summativen *Assessment of Learning* und dem formativen *Assessment for Learning* (1,3–5).

Der Begriff *Test-enhanced learning* beschreibt den damit eng zusammenhängenden psychologischen Effekt, dass Lernende Inhalte in der Zukunft besser abrufen können, wenn sie bereits im Rahmen einer Prüfung bzw. eines Tests abgefragt wurden, als wenn sie nur wiederholt ohne Prüfungen gelernt wurden (6,7). Der Effekt lässt sich durch Feedback verstärken (4,6–8).

Die ideale Prüfung sollte häufig stattfinden, um kontinuierliches Lernen anzuregen, eher Kompetenzen als reines Fachwissen prüfen und detailliertes Feedback zu Stärken und Schwächen geben (3). Um möglichst Kompetenzen statt reines Fachwissen zu prüfen, wurden neue Prüfungsformate entwickelt, z.B. Prosa-basierte schriftliche Klausuren, praktische Stationsprüfungen (sog. OSCEs), arbeitsplatzbasierte Prüfungen, Portfolios und viele andere (2,3,9). Wegen des hohen Aufwandes dieser Prüfungsformate eignen sich diese aber nur begrenzt für in kurzen Zeitabständen stattfindende Prüfungen. Diese wären aber notwendig, um die Mechanismen des Test-enhanced learning auszunutzen. Angesichts dessen werden „klassische“ MC-Fragen-basierte Klausuren weiterhin vielfach im Medizinstudium weltweit eingesetzt, auch an der Charité sind die meisten Prüfungen MC-basiert (9–11).

1.2 Verwendung von MC-Fragen

1.2.1 Psychometrische Parameter

In der Literatur ist „Item“ der gängige Begriff für eine Prüfungsaufgabe, in unserem Fall eine einzelne MC-Frage. Diese Arbeit konzentriert sich auf die zwei wichtigsten psychometrischen Itemparameter. Sie werden jeweils nach der Prüfung berechnet:
1) Die *„Schwierigkeit“* bzw. der „*Schwierigkeitskoeffizient*“ ist definiert als die durchschnittliche Punktzahl der Teilnehmenden geteilt durch die maximal erreichbare Punktzahl des Items. (Bei MC-Fragen mit nur einer richtigen Antwort wie in unserem Fall stimmt der Wert mit der relativen Anzahl der Studierenden, die die Frage richtig beantwortet haben, überein.) Er kann Werte zwischen 0 und 1 annehmen, wobei höhere Werte ein leichter zu beantwortendes Item beschreiben. In der klassischen Testtheorie, die vor allem eine gute Differenzierung zwischen Studierenden zum Ziel hat, sollen Items möglichst eine Schwierigkeit zwischen 0,4 und 0,8 haben, zu leichte Items dienen nicht der Differenzierung. In kompetenzbasierten Curricula und Prüfungen dürfen Items auch leichter sein, wenn sie basale Kenntnisse und Fähigkeiten prüfen. (10,12–14)
2) Die „*Trennschärfe“* bzw der *„Trennschärfekoeffizient“* beschreibt die Fähigkeit eines Items, zwischen Studierenden mit unterschiedlichem Wissensstand zu differenzieren. Da man den Wissensstand in seiner Gesamtheit schlecht objektiv messen und korrelieren kann, behilft man sich mit der Gesamtnote der Prüfung als Marker. Eine Frage mit hoher Trennschärfe wird von Studierenden mit insgesamt guter Note überdurchschnittlich häufiger richtig beantwortet als von Studierenden mit schlechter Gesamtnote. Es existieren verschiedene Berechnungsmethoden: Eine einfache Möglichkeit besteht darin, für jedes Item die Differenz *D* der durchschnittlichen Punktzahl der besten und der schlechtesten 33% (je nach Quelle gelegentlich auch 27%) der Studierenden zu bilden, den sog. „Index of Discrimination“. Die alternative Methode besteht in der Berechnung des korrigierten Korrelationskoeffizienten *r‘* nach Pearson für die in der Aufgabe erreichten Punktzahlen mit der Summe der Punkte in allen anderen Fragen. *r‘* liegt zwischen -1 und 1, nach der klassischen Testtheorie gelten in MC-Prüfungen Werte > 0,3 als gut, zwischen 0,2 und 0,3 als akzeptabel. Insb. sehr einfache Aufgaben haben niedrige Trennschärfen nahe 0. Werte < 0 beschreiben eine negative Korrelation, d.h. die Frage wird von insgesamt schlecht abschneidenden Studierenden häufiger richtig beantwortet als von guten. Items

mit diesem sog. paradoxen Antwortmuster müssen genau auf eventuelle inhaltliche Fehler geprüft werden. Allgemein ist *r‘* als Berechnungsmethode der Trennschärfe vorzuziehen, da es mathematisch exakter ist, insb. bei sehr leichten Items. (10,12,13)

1.2.2 Vor- und Nachteile von MC-Fragen

Der wahrscheinlich wichtigste Grund für die Beliebtheit von MC-Fragen ist der insgesamt geringe Aufwand für Prüfende: Die Items lassen sich schnell erstellen (insbesondere solche von geringer Qualität, s.u.), die Prüfung ist logistisch weniger aufwendig als z.B. ein OSCE, lässt sich leicht und maschinell statistisch auswerten und es sind wenige Aufsichtspersonen notwendig. Darüber hinaus ist die Prüfung insbesondere bei großen Kohorten mit heterogenen Prüfenden und Teilnehmenden objektiver und reliabler als z.B. mündliche oder Prosa-basierte Prüfungen (8) und eignet sich damit gut als Format für große Kohorten. Nahezu alle großen, nationalen Lizenzierungsprüfungen wie z.B. das deutsche Staatsexamen und das USMLE sind zumindest teilweise MC-basiert (15–17). Entsprechend ist es auch im Interesse der Fakultäten, ihre Studierenden im Prüfungsformat MC zu „trainieren“, um möglichst gute Ergebnisse in nationalen Vergleich zu erreichen.

Der Hauptnachteil besteht in der Art des abgeprüften Wissens. Nach Miller (18) werden vier Stufen klinischen Wissens unterschieden: „Weiß“, „Weiß wie“ (er/sie das Wissen anwendet), „Zeigt“ (wie er/sie das Wissen anwendet), „Tut“ (wendet das Wissen praktisch an). Normale MC-Fragen können nur Stufe 1 des klinischen Wissens abdecken, komplizierte, z.B. Vignetten-basierte MC-Fragen können auch Stufe 2 abdecken, sind aber aufwendiger zu konstruieren (19). Die darüberhinausgehenden, oben genannten und in modernen Curricula verlangten Kompetenzen lassen sich nicht überprüfen. MC-Fragen können also als alleiniges Prüfungsformat nicht ausreichen. Ein weiterer Nachteil besteht in den verwendeten Falschantworten, den sog. Distraktoren. Merken sich Prüflinge die Distraktoren als richtige Antwort, wird falsches Wissen über den oben beschriebenen Mechanismus des Test-enhanced learning verinnerlicht. Es ist didaktisch essenziell, dass dieser Retention falscher Inhalte durch individuelles Feedback über die richtigen Antworten entgegengewirkt wird (8). Gleichzeitig sollten Distraktoren sinnvoll gewählt sein, damit sowohl eine Antwort eindeutig richtig ist, diese aber nicht im Ausschlussverfahren ohne eigentliche Kenntnis des Inhaltes rückgeschlossen werden kann. Dabei gibt es eine Viel-

zahl an Fallstricken und entsprechende Anleitungen zur Erstellung hochwertiger MC-Fragen, was deren Erstellung oft aufwendiger macht als von vielen Prüfenden angenommen (19,20).

1.3 Veröffentlichung und Wiederverwendung von Prüfungsitems

Durch die hohe Anzahle benötigter Items und den hohen Aufwand der Erstellung immer neuer Items, insbesondere, wenn diese einen gewissen formellen Mindeststandard erfüllen sollen, sind viele Fakultäten auf die Wiederverwendung von Fragen angewiesen (10,11,21). Dadurch entsteht aber auch die Möglichkeit der Weitergabe der Items an nachfolgende Studierende, die die Prüfung noch absolvieren müssen. Diese Praxis ist in medizinischen Studiengängen weit verbreitet und wird in der Regel nicht als Betrug wahrgenommen (21). Bei vorheriger Kenntnis der Antwort fällt entsprechend insbesondere bei aufwendig konstruierten, vignettenbasierten Fragen viel Denkleistung weg, die betroffenen Items werden leichter und die Reliabilität der Prüfung könnte sinken.
Für kompetenzbasierte Curricula ergibt sich entsprechend ein Dilemma: Didaktisch ist es notwendig, Studierenden Feedback über ihre richtigen und falschen Klausurantworten zu geben. Dafür muss Studierenden nach der Prüfung Zugriff auf die verwendeten Items und die Antworten gewährt werden. Das wiederum könnte zu einer vermehrten Weitergabe an nachfolgende Generationen von Studierenden führen, welche mit weniger Lernaufwand bestehen könnten. Möglicherweise sorgt man durch die Veröffentlichung der Items also für eine schlechte Lernsteuerung und senkt die Reliabilität der Prüfung.
Diesen potenziell großen Nachteilen der Veröffentlichung und Wiederverwendung von Items stehen neben dem verbesserten Feedback noch weitere Vorteile gegenüber. Insbesondere bei Prüfungsergebnissen, die die weitere akademische Laufbahn beeinflussen (Nichtbestehen, Erfüllung der Voraussetzungen von Stipendien, etc.), fordern Studierende zurecht mehr Transparenz, welche sich nur durch eine Einsichtnahme der Prüfungsunterlagen herstellen lässt. Gelegentlich gibt es unter Umständen auch juristische Bestimmungen, die das Recht auf Einsichtnahme festlegen (22), sodass Fakultäten durch die generelle Veröffentlichung Gerichtsprozessen vorbeugen könnten. Ein weiterer Vorteil ist möglicherweise verminderte Prüfungsangst, wenn eine Möglichkeit zur Vorbereitung mit Original-Items besteht (23–25).

1.4 Forschungsstand

Die Effekte der Veröffentlichung und Wiederverwendung von Prüfungsitems in medizinischen Studiengängen sind angesichts des oben beschriebenen Dilemmas und der weitverbreitenden Praxis der Wiederverwendung von Items erstaunlich schlecht untersucht (10,22,26).

Es gibt wenige Studien, die sich mit der Wiederverwendung von MC-Fragen in medizinischen Fakultäten beschäftigen. Joncas et al. (2018) untersuchten 1.629 Items, die innerhalb des fünfjährigen Studienzeitraums bei jeder Wiederholung leichter wurden und an Trennschärfe einbüßten. Jedoch war der Effekt eher klein und die Autoren haben keine Angaben dazu gemacht, welche Geheimhaltungsmaßnahmen der verwendeten Fragen bestanden (11). Wagner-Menghin et al. (2013) konnten bei wiederholten Prüfungen mit jeweils einem Anteil wiederverwendeter Fragen keinen Einfluss auf Item-Schwierigkeit und Gesamtnote feststellen, machten jedoch ebenfalls keine Angaben zu Geheimhaltungsmaßnahmen (27). Herskovic (1999) konnte keine Veränderung der psychometrischen Parameter von 197 in Folgeexamina wiederverwendeten Items feststellen, nachdem diese mündlich mit den Prüflingen besprochen wurden (28). Wood (2009) konnte nachweisen, dass Prüflinge, die eine Prüfung erneut antreten mussten, bei Items, die sie bereits kannten, nicht besser abschnitten als bei unbekannten (29).

Yang et al. (2018) stellten keine signifikanten Veränderungen der Item-Parameter, Prüfungsnoten oder Bestehensquoten im südkoreanischen Äquivalent zum deutschen Staatsexamen fest, nachdem die verwendeten MC-Fragen und Antworten veröffentlicht wurden. Diese wurden nicht wiederverwendet, jedoch war vorher die Vorbereitung mit Fragenkatalogen aus Gedächtnisprotokollen weit verbreitet (17).

Auch außerhalb der Medizin gibt es nur wenige Untersuchungen zu den Auswirkungen der Wiederverwendung und Veröffentlichung von Items, die Ergebnisse sind auch hier nicht eindeutig (22,26,30).

1.5 Ausgangsposition und Fragestellung dieser Studie

Zur Vermeidung der Bekanntwerdung der Prüfungsitems bestanden an der Charité-Universitätsmedizin Berlin in der Vergangenheit strenge Geheimhaltungsmaßnahmen der in Modul- bzw. Semesterabschlussklausuren des Modellstudiengangs Humanmedizin verwendeten MC-Fragen: Studierende mussten ihr Testheft abgeben, die Antworten wurden

nicht veröffentlicht und die Einsichtnahme zwecks der Formulierung möglicher Einsprüche war nur in eingeschränkten Zeiträumen, nach Terminvereinbarung und unter Aufsicht möglich. Beginnend mit den Prüfungen zum Abschluss des Sommersemesters 2017 entschied sich der Prüfungsausschuss, diese Geheimhaltungsmaßnahmen deutlich zu lockern: Die Mitnahme der Testhefte war ab sofort gestattet und nach der Prüfung wurden die korrekten Antworten online veröffentlicht. Obwohl die Verbreitung weiterhin formell untersagt war, kam es in der Folge trotzdem zu einer de facto-Veröffentlichung der Fragentexte und Antworten über filesharing-Plattformen, welche auch für Studierende einsehbar waren, die die Prüfung noch absolvieren mussten und die mit der zumindest teilweisen Wiederverwendung der einsehbaren Fragen rechnen konnten. (10)
Die Gründe für diesen Entschluss waren mannigfaltig und wurden teilweise auch schon in den darüberliegenden Abschnitten diskutiert: Erhöhung der Transparenz, Verringerung des Arbeitsaufwandes der Mitarbeiter des Prüfungsbereichs, Verminderung der Prüfungsangst, Möglichkeit zum verbesserten Feedback, sowie die Annahme, dass große Teile der Prüfungen bereits durch Gedächtnisprotokolle der Studierenden bekannt waren (21). Außerdem wurde angenommen, dass es aufgrund der Vielzahl der bereits in der Datenbank vorhandenen Items (20.000 bzw. im Mittel 2.000 pro Semester) erstens sehr lange dauern würde, bis diese alle veröffentlicht werden, zweitens, dass sich Studierende diese Vielzahl an Items gar nicht vollständig merken können und drittens, die vorhanden Items bereits die wichtigsten Inhalte der unterrichteten Inhalte abdecken und somit den im Rahmen von kompetenzbasierter medizinischer Lehre geforderten Mindeststandard abdecken (14).

Angesichts der widersprüchlichen Studienlage ergibt sich aus dieser Konstellation eine einmalige Chance, den Einfluss der Wiederverwendung und Veröffentlichung von Items auf psychometrische Parameter der Items und auf die Prüfungsergebnisse auszuwerten. Die Entscheidung zur Veröffentlichung bei gleichzeitig unveränderter Praxis der Wiederverwendung erlaubt sowohl den Vergleich zwischen veröffentlichen und nicht veröffentlichten Items als auch den retrospektiven Vergleich der psychometrischen Parameter bei Erst- und Wiederverwendung in einem quasi-experimentellen Setting (10).

Für die vorliegende Arbeite ergeben sich drei Hauptfragestellungen:

1. Gibt es Unterschiede in Schwierigkeit und Trennschärfe zwischen neuen, bisher nicht veröffentlichten und wiederverwendeten sowie veröffentlichten und wiederverwendeten Items?
2. Wie verändern sich Schwierigkeit und Trennschärfe bei Wiederverwendung eines Items, abhängig davon, ob es bisher nicht oder bereits veröffentlicht wurde?
3. Wie verändern sich die gesamte, mittlere Schwierigkeit und Trennschärfe sowie Prüfungsnoten, Bestehensquoten und der Anteil der von der Gleitklausel (s.u.) betroffenen Studierenden bei steigendem Anteil veröffentlichter und wiederverwendeter Items?

Die Untersuchung der Itemparameter wurde bereits in der zu dieser Dissertation gehörigen Publikation veröffentlicht. Die Analyse von Prüfungsnoten, Bestehensquoten und dem Anteil der von der Gleitklausel betroffenen Studierenden wurde bisher nicht publiziert.

2 Methodik

2.1 Rahmenbedingungen

Die Charité-Universitätsmedizin Berlin ist eine der größten medizinischen Fakultäten in Europa mit zum Zeitpunkt der Studie ca. 4.500 immatrikulierten Studierenden. Das Studium ist in Semester und Module gegliedert, jeweils am Ende der Semester 1 bis 10 werden die jeweiligen Module durch eine Prüfung abgeschlossen. Dabei handelt es sich um eine schriftliche Prüfung mit zum Studienzeitpunkt 40 bis 120 Einzelantwort-MC-Items. Es gibt zwei Prüfungstermine mit unterschiedlichen Klausurversionen pro Semester, jeweils am Anfang und Ende der Semesterferien.
Wie das Studium sind auch die Prüfungsinhalte interdisziplinär mit verschiedenen, enthaltenen Grundlagen- und klinischen Fächern. Alle Klausuren werden standardisiert erstellt: Die zu prüfenden Lernziele werden anhand eines Blueprints zufällig ausgewählt. Für bis zu 80% dieser Lernziele werden zufällig passende Items ausgewählt, die bereits in zurückliegenden Semestern entwickelt und verwendet wurden. Zum Studienzeitpunkt enthielt diese Datenbank über 20.000 MC-Fragen. Die Fragen für die verbleibenden 20% werden von Fakultätsmitgliedern neu erstellt. Um eine hohe inhaltliche Qualität sicherzustellen, werden die neuen Fragen sowie die finalen Klausuren nochmals von mindestens zwei inhaltlich verantwortlichen Fakultätsmitgliedern sowie einem ärztlichen Mitarbeiter bzw. Mitarbeiterin der Prüfungsabteilung korrekturgelesen. Nach der Prüfung können Studierende Einspruch gegen potenziell fehlerhafte Items einreichen und der Prüfungsausschuss ändert ggf. deren Wertung.
Klausuren an der Charité werden nach dem allgemeinen deutschen Notensystem von 1 bis 5 bewertet, wobei 1,0 die beste Note mit 100% richtigen Antworten und 4,49 die schlechteste Note darstellt, mit der man noch besteht. Note 5 ist gleichbedeutend mit dem Nicht-Bestehen der Prüfung. Die Bestehensgrenze beträgt regulär 60 % richtig beantworteter Items. (10)

2.1.1 Gleitklausel

Gleichzeitig darf die Bestehensgrenze nur 22 % unterhalb der durchschnittlichen Prüfungsleistung der erstmals teilnehmenden Studierenden liegen. Diese sog. „Gleitklausel“ sorgt in der Praxis dafür, dass die Bestehensgrenze unter 60 % sinkt, sobald die durchschnittliche Prüfungsleistung der erstmals teilnehmenden Studierenden unterhalb von 77

% liegt. Zur Verdeutlichung: liegt die durchschnittliche Prüfungsleistung der Erstteilnehmenden bei 71%, so beträgt die Bestehensgrenze: 71 % - (71 % x 0,22) = 55,38 %. Die Gleitklausel, wie sie auch im Rahmen des 2. Staatsexamens nach ÄApprO verwendet wird, dient dazu, ungewöhnlich hohe Durchfallquoten zu verhindern. (31,32)

2.2 Untersuchte Parameter

Wir untersuchten alle 5 Klausurperioden vom Sommersemester 2017 bis zum Sommersemester 2019, wobei es jeweils 2 Klausuren mit unterschiedlichen Fragen je Semester gab. Die Klausuren des Sommersemesters 2017 waren die ersten, die im Anschluss an die Prüfungen veröffentlicht wurden, sodass diese als Baseline ohne wiederverwendete, bereits veröffentlichte Items dienten. Die nachfolgenden Klausuren enthielten wiederverwendete, bereits veröffentliche Items mit steigendem Anteil.
Für alle inkludierten Klausuren wurden retrospektiv die Anzahl der teilnehmenden Studierenden sowie die Schwierigkeitskoeffizienten und die Trennschärfe als Korrelationskoeffizient *r‘* der verwendeten Items bestimmt (s. Abschnitt 1.2.1 der Einleitung). Für wiederverwendete Items wurde außerdem der Schwierigkeits- und Trennschärfekoeffizient bei Erstverwendung bestimmt.
Items wurden ausgeschlossen, wenn sie im Review nach der Klausur in der Wertung modifiziert wurden.
Die Items wurden wie folgt gruppiert:

1. *Neu:* Erstmals in dieser Klausur eingesetzt
2. *Wiederverwendet, nicht veröffentlicht:* Bereits in einer vorherigen Klausur eingesetzt, allerdings zuletzt vor dem Sommersemester 2017.
3. *Wiederverwendet, veröffentlicht:* Bereits in einer vorherigen Klausur eingesetzt, zuletzt im Sommersemester 2017 oder später.

Dabei bestand die Möglichkeit, dass ein Item im Studienzeitraum mehrfach verwendet werden konnte, mit jeweils unterschiedlichen Itemparametern. Die Items wurden für die jeweilige Verwendung entsprechend der oben beschriebenen Regeln gruppiert. (10)

Zusätzlich zu den im Rahmen der Publikation (10) veröffentlichten Daten wurden die Prüfungsnoten der teilnehmenden Studierenden ausgewertet. Die Bestehensquote entspricht dabei der Relation von Studierenden mit Note 4 oder besser und allen teilnehmenden Studierenden. Um Verzerrungen durch die Gleitklausel besser nachvollziehen zu können, wurde pro Semester die Anzahl der von der Gleitklausel betroffenen Studierenden berechnet.

2.3 Analyse

Entsprechend der Fragestellungen wurden folgende Analysen vorgenommen:

1. Um die Unterschiede zwischen den oben beschriebenen Item-Gruppen zu analysieren, wurden Schwierigkeits- und Trennschärfe-Koeffizient als jeweils abhängige Variable mit einer ANOVA untersucht, wobei die Item-Gruppe als unabhängige Variable diente. (10)
2. Um den relativen Effekt der Veröffentlichung auf Schwierigkeit und Trennschärfe bei Wiederverwendung eines Items zu analysieren, wurde für wiederverwendete Items die Differenz der Schwierigkeits- und Trennschärfe zwischen Nutzung im Studienzeitraum und bei erstmaliger Nutzung gebildet. Diese Differenz diente als jeweils unabhängige Variable und wurde jeweils mittels zweier t-Tests zwischen den Item-Gruppen verglichen. (10)
3. Um die Veränderung der Prüfungsnoten, Bestehensquoten, Verwendung der Gleitklausel, Schwierigkeits- und Trennschärfekoeffizienten zu analysieren, wurden diese als jeweils abhängige Variable mit einer ANOVA untersucht, wobei die Klausurperiode als unabhängige Variable diente.

Dabei galten folgende Grundsätze:

Bei signifikantem ANOVA-Ergebnis dienten Bonferroni-Post-Hoc-Tests zum individuellen Vergleich zwischen den Gruppen. Die Effektstärke für die ANOVA wurde als η_p^2 berechnet. Wir folgten Cohens Definition für kleine ($\eta_p^2 = 0.01$), mittlere ($\eta_p^2 = 0.06$) und große ($\eta_p^2 = 0.14$) Effekte. (33) Die Effektstärke für t-Tests wurde als Cohen‘s *d* berechnet, mit definierten Größenordnungen für kleine ($d = 0.2$), mittlere ($d = 0.5$), and große ($d = 0.8$) Effekte. (10)

Alle Berechnungen wurden mittels *IBM SPSS 25* und *29* vorgenommen. Die verwendeten Grafiken wurden mit *GraphPad Prism* und *diagrammeditor.de* erstellt. (34–36)

3. Ergebnisse

Wir untersuchten insgesamt 199 Klausuren mit 23.507 Teilnehmenden. Von 10.600 Items verwendeten Items im Studienzeitraum konnten wir 10.148 einschließen. Der Prozentsatz der Prüfungswiederholenden betrug 4,8%. *Abbildung 1* gibt einen Überblick über die Anzahl der eingeschlossenen Items.

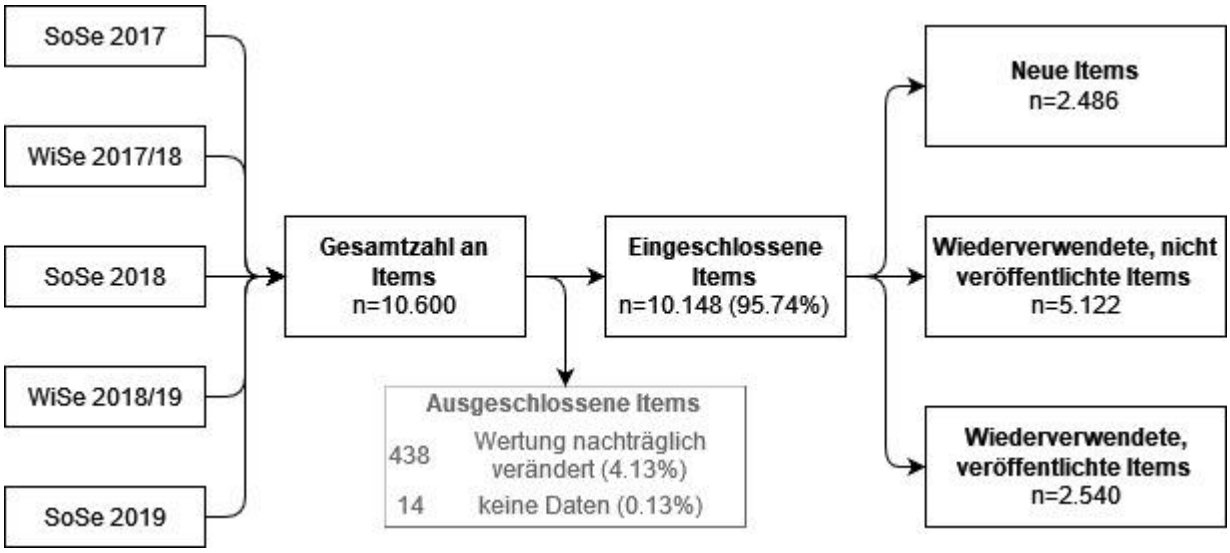

Abbildung 1: Übersicht über eingeschlossene Items. Quellenangabe: Appelhaus et al., 2023 (10)

3.1 Unterschiede zwischen den Itemgruppen

Neue Fragen waren mit einem Schwierigkeitskoeffizienten von M = 0,66 am schwersten zu beantworten, gefolgt von wiederverwendeten, nicht veröffentlichten (M = 0,71) und wiederverwendeten, veröffentlichten Fragen, welche am leichtesten zu beantworten waren (M = 0,83). Die ANOVA ergab einen signifikanten Unterschied von mittlerer Effektstärke, $F(2, 10.145) = 483,38$, $P < 0,001$, $\eta_p^2 = 0,087$. Bonferroni Post Hoc Tests ergaben signifikante Unterschiede zwischen allen Gruppen. (10)

Neue (M = 0,20) und wiederverwendete, nicht veröffentlichte Fragen (M = 0,21) hatten einen niedrigeren Schwierigkeitskoeffizienten als wiederverwendete, veröffentlichte Fragen (M = 0,25). Die ANOVA ergab einen signifikanten Unterschied von niedriger bis vernachlässigbarer Effektstärke, $F(2, 9.736) = 49,50$, $P < 0,001$, $\eta_p^2 = 0,008$. Bonferroni Post-hoc-Tests ergaben signifikante Unterschiede nur zwischen wiederverwendeten, veröffentlichten Fragen und den anderen beiden Itemgruppen. (10)

Abbildung 2 gibt einen Überblick über die Ergebnisse.

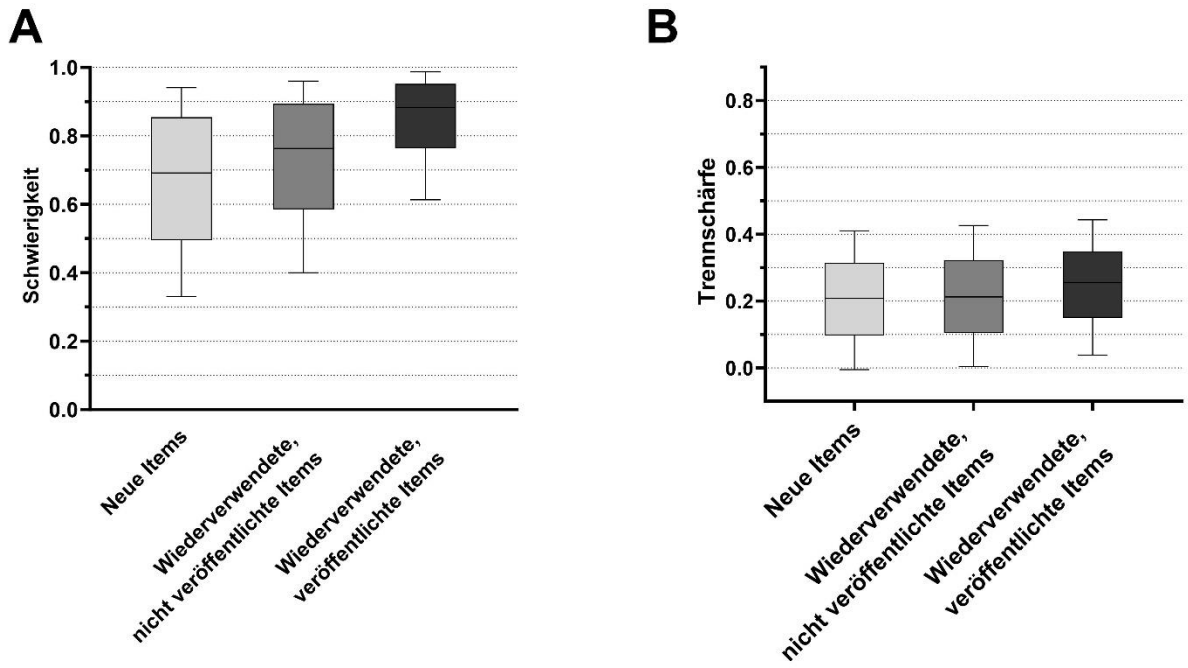

Abbildung 2: Schwierigkeitskoeffizienten (A) und Trennschärfe (B) der einzelnen Itemgruppen. Quellenangabe: Appelhaus et al., 2023 (10)

3.2 Veränderung zur Erstverwendung

Verglichen mit der erstmaligen Verwendung haben wiederverwendete, nicht veröffentlichte Items einen um M = -0,01 niedrigeren und wiederverwendete, veröffentlichte Items einen um M = 0,11 höheren Schwierigkeitskoeffizienten. Der t-Test ergab einen signifikanten Unterschied zwischen den Itemgruppen von mittlerer bis hoher Effektstärke, $t(4699{,}79) = -29{,}69$, $P < 0{,}001$, $d = 0{,}74$. (10)

Die Trennschärfe erhöhte sich bei wiederverwendeten, nicht veröffentlichten Items um M = 0,03 und bei wiederverwendeten, veröffentlichten Items um M = 0,07. Der t-Test ergab einen signifikanten Unterschied von niedriger bis vernachlässigbarer Effektstärke, $t(4274{,}07) = -7{,}22$, $P < 0{,}001$, $d = 0{,}19$.

Abbildung 3 gibt einen Überblick über die Ergebnisse.

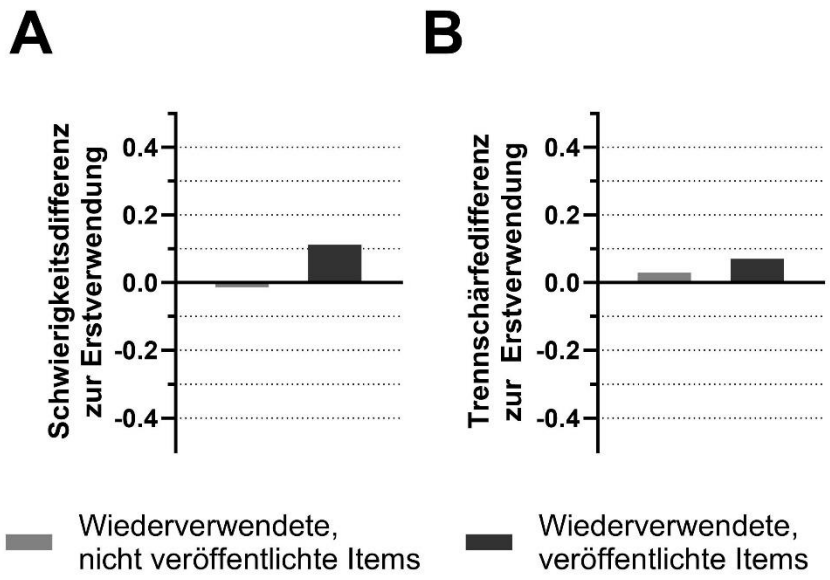

Abbildung 3: Veränderung der Schwierigkeitskoeffizienten (A) und Trennschärfe (B) wiederverwendeter, nicht veröffentlichter und veröffentlichter Items im Vergleich zur ersten Verwendung. Quellenangabe: eigene Abbildung.

3.3 Veränderung der Itemparameter und Noten über die Zeit

Der Anteil wiederverwendeter, veröffentlichter Items stieg im Studienzeitraum kontinuierlich von 0 % im Sommersemester 2017 auf 48,4% im Sommersemester 2019. ((10), *Abbildung 4a*)

Die Prüfungsnoten der bestandenen Studierenden lagen im Mittel zwischen M = 2,64 im Sommersemester 2017 und M = 2,41 im Wintersemester 2018. Die ANOVA ergab signifikante Unterschiede bei niedriger Effektstärke, $F(4, 22.124) = 53,89$, $P < 0,001$, $\eta_p^2 = 0,010$. Die Bonferroni Post-hoc-Tests wiesen signifikante Unterschiede allerdings ausschließlich zwischen Sommersemester 2017 und Wintersemester 2018 nach. (*Abbildung 4b*)

Die Bestehensquoten lagen zwischen 93,8% im Sommersemester 2017 und 94,9% im Wintersemester 2018 ohne signifikante Unterschiede, $F(4, 23.501) = 1,73$, $P = 0,141$. Der Anteil der von der Gleitklausel betroffenen Prüfungsteilnehmenden sank im Studienverlauf von 39,7 % im Sommersemester 2017 auf 26,9 % im Sommersemester 2019, $F(4, 23.501) = 81,50$, $P < 0,001$, $\eta_p^2 = 0,014$. (*Abbildung 4c)*

Die Schwierigkeitskoeffizienten lagen im Mittel zwischen M = 0,70 im Sommersemester 2017 und M = 0,76 im Wintersemester 2018. Die ANOVA ergab signifikante Unterschiede bei niedriger Effektstärke, $F(4, 10.143) = 27,03$, $P < 0,001$, $\eta_p^2 = 0,011$. Die Trennschärfe

lag im Mittel zwischen *M* = 0,21 im Sommersemester 2017 und *M* = 0,23 im Wintersemester 2018. Die ANOVA ergab signifikante Unterschiede bei vernachlässigbarer Effektstärke, $F(4, 9.734) = 4,25$, $P = 0,002$, $\eta_p^2 = 0,002$ (*Abbildung 4d)* (10). Aus Übersichtsgründen sind die Ergebnisse doppelt aufbereitet, als *Tabelle 1* und *Abbildung 4*.

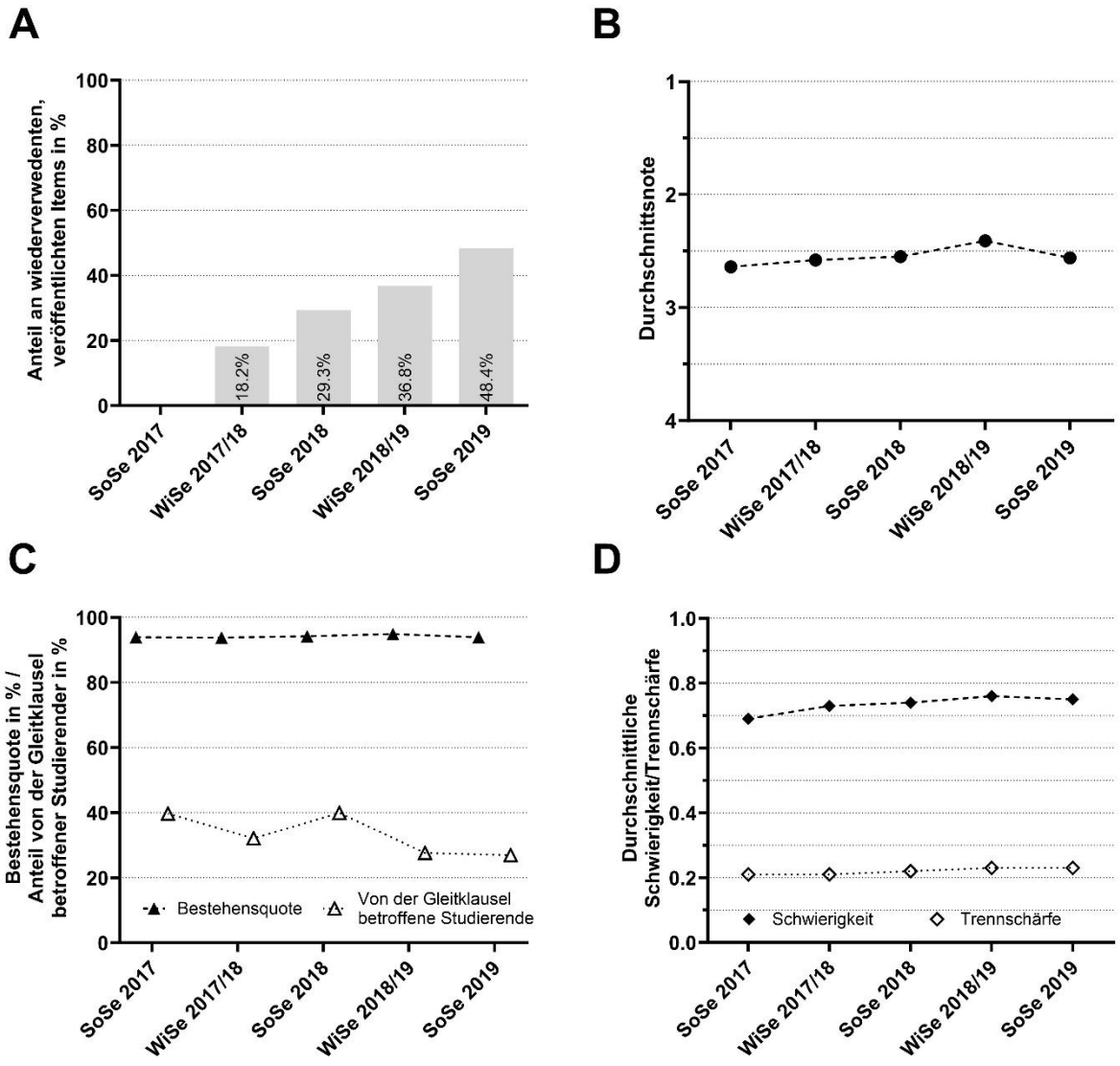

Abbildung 4: Veränderung des Anteils der wiederverwendeten, veröffentlichten Items (A) der mittleren Note bestandener Studierender (B) der Bestehensquote und des Anteils der von der Gleitklausel betroffenen Studierenden (C) sowie der mittleren Schwierigkeit und Trennschärfe im Studienzeitraum (D). Quellenangabe: Diagramme A und D aus Appelhaus et al., 2023 (10)

Tabelle 1: Detaillierte Übersicht über die Prüfungsnoten, Bestehensquoten, Verwendung der Gleitklausel, Schwierigkeits- und Trennschärfekoeffizienten im Studienzeitraum. Quellenangabe: Zeilen 5 und 6 aus Appelhaus et al., 2023 (10):

	SoSe 2017	WiSe 2017/18	SoSe 2018	WiSe 2018/19	SoSe 2019	ANOVA	Bonferroni Post-hoc-Tests
Note bestandener Teilnehmer *(SD)*	2,64 (0,76)	2,58 (0,79)	2,55 (0,78)	2,41 (0,80)	2,56 (0,77)	$F(4, 22.124) = 53,89$, $P < 0,001$, $\eta_p^2 = 0,010$	signifikante Unterschiede nur jeweils zwischen SoSe 2017 und WiSe 2018/19 zu allen anderen
Bestehensquote *(SD)*	93,9 % (0,24)	93,8 % (0,24)	94,2 % (0,23)	94,9 % (0,22)	93,9 % (0,24)	$F(4, 23.501) = 1,73$, $P = 0,141$	—
Von der Gleitklausel betroffene Teilnehmer *(SD)*	39,7 % (0,49)	32,1 % (0,49)	39,9 % (0,49)	27,6 % (0,45)	26,9 % (0,44)	$F(4, 23.501) = 81,50$, $P < 0,001$, $\eta_p^2 = 0,014$	signifikante Unterschiede zwischen allen Gruppen außer zwischen SoSe 2017 und SoSe 2018 sowie zwischen WiSe 2018/19 und SoSe 2019
Schwierigkeits-koeffizient *(SD)*	0,70 (0,23)	0,73 (0,21)	0,74 (0,21)	0,76 (0,21)	0,75 (0,21)	$F(4, 10.143) = 27,03$, $P < 0,001$, $\eta_p^2 = 0,011$	Signifikante Unterschiede zwischen allen Gruppen außer zwischen SoSe 2018, WiSe 2018/19 und SoSe 2019 sowie zwischen WiSe 2017/18 und SoSe 2018
Trennschärfe-koeffizient (SD)	0,21 (0,18)	0,21 (0,17)	0,22 (0,16)	0,23 (0,18)	0,23 (0,18)	$F(4, 9.734) = 4,25$, $P = 0,002$, $\eta_p^2 = 0,002$	Signifikante Unterschiede zwischen allen Gruppen außer zwischen SoSe 2017 und WiSe 2018/19 sowie zwischen SoSe 2017 und SoSe 2019

4. Diskussion

4.1 Kurze Zusammenfassung der Ergebnisse

Wir konnten nachweisen, dass wiederverwendete, veröffentlichte Items in MC-Prüfungen leichter zu beantworten sind als wiederverwendete, nicht veröffentlichte Items und neue Items. Verglichen mit der jeweils ersten Verwendung waren wiederverwendete Items leichter zu beantworten, während sich die Schwierigkeit wiederverwendeter, veröffentlichter Items nicht signifikant von der ersten Verwendung unterschied, was darauf hinweist, dass die Veröffentlichung der maßgebende Faktor für die sinkende Schwierigkeit bei Wiederverwendung ist. Die Trennschärfe war gering und durch Wiederverwendung und Veröffentlichung eher positiv beeinflusst. (10)
Der Effekt auf die durchschnittliche Gesamtnote war gering, zwischen dem besten und dem schlechtesten Semester im Studienzeitraum unterschied sie sich nur um 0,23 auf einer Skala von 1 bis 4,49. Es konnte kein Effekt auf die Bestehensquote nachgewiesen werden, allerdings ging der Anteil der von der Gleitklausel betroffenen Teilnehmenden zurück.

4.2 Interpretation der Ergebnisse

Die oben beschriebene, unterschiedliche Schwierigkeit mit deutlich einfacher zu beantwortenden Fragen nach Veröffentlichung bestätigt im Wesentlichen die Vermutung von Kritiker*innen und lässt vermuten, dass die von der Charité gewählte Art der Veröffentlichung das Teilen der Prüfungsinhalte mit nachfolgenden Kohorten deutlich einfacher macht. Überraschend war, dass wiederverwendete, nicht veröffentlichte Fragen keinerlei Unterschiede zwischen erster Verwendung und wiederholter Verwendung im Studienzeitraum aufwiesen. Das zeigt, dass die vorher bestehenden Geheimhaltungsmaßnahmen wahrscheinlich sehr erfolgreich waren und die Fragen nicht in relevanter Anzahl und in verwertbarer Form zur Prüfungsvorbereitung zur Verfügung standen.
Die beobachtete, leicht steigende Trennschärfe ist ebenfalls ein überraschendes Ergebnis und erklärt sich möglicherweise daraus, dass es Studierenden mit insgesamt besserer Note leichter fallen könnte, sich die veröffentlichten Fragen zu merken. Damit lässt sich zeigen, dass die Veröffentlichung der Fragen wahrscheinlich keine Auswirkungen auf das Ranking der Studierenden hat. (10)

Auf den ersten Blick ebenfalls überraschend ist der geringe Effekt auf Prüfungsnoten und Bestehensquoten. Eine Note von 1,0 entspricht 100% richtigen Antworten, eine Note von 4,49 entspricht 60% richtigen Antworten. Für einen ganzen Notenpunkt Verbesserung muss man also den Anteil richtiger Antworten um absolut ca. 11,4% erhöhen. Bei einem Anteil von fast 50% wiederverwendeten, veröffentlichten Items pro Klausur im Wintersemester 2019 und einem Unterschied des Schwierigkeitskoeffizienten von $d > 0{,}1$ zu den anderen Itemgruppen sollte die Gesamtnote also rechnerisch um ca. 0,5 Punkte besser werden. Beobachtet wurden aber nur 0,12 Notenpunkte Verbesserung. Die Bestehensquote war nicht signifikant verändert. Dieser Beobachtung liegen wahrscheinlich zwei kombinierte Effekte zugrunde. Erstens, und hauptsächlich, lässt sich das Ergebnis durch die Gleitklausel erklären. Die Gleitklausel dient dazu, hohe Durchfallquoten in schlecht ausgefallenen Prüfungen zu verhindern, indem sie die zum Bestehen und für bestimmte Noten notwendige Punktzahl nach unten korrigiert, wie im Abschnitt Methodik dieser Arbeit beschrieben. Durch die zunehmende Verwendung wiederverwendeter, veröffentlichter Items mit höherem Schwierigkeitskoeffizienten kommt es wie beobachtet seltener zur Anwendung der Gleitklausel, sodass die Bestehensquote gleichbleibt, während die für bestimmte Noten notwendige Punktzahl angehoben wird. Zweitens kommt es nach einem Peak im Wintersemester 2018 im Sommersemester 2019 bereits wieder zu einem Abfall der Schwierigkeitskoeffizienten und Bestehensquote, trotz weiter steigendem Anteil wiederverwendeter, veröffentlichter Fragen. Dabei könnte es sich um einen statistischen Ausreißer handeln, wobei die große Anzahl der verwendeten Fragen dies unwahrscheinlich macht, zumal man eigentlich einen Anstieg erwarten würde. Wie bereits im Abschnitt 1.5 der Einleitung hypothetisiert, lässt sich der Effekt eventuell aus der Kombination erklären, die zu diesem Zeitpunkt am Ende des Studienzeitraumes insgesamt ca. 8.000 und pro Prüfung bis zu 480 bereits veröffentlichten Items schlicht nicht mehr zu memorisieren waren und der Effekt der Veröffentlichung auf die Itemparameter ein Plateau erreicht hat. Um diese Hypothese zu verifizieren wäre eine Fortsetzung dieser Studie über die nachfolgenden Semester notwendig.

4.3 Einbettung der Ergebnisse in den bisherigen Forschungsstand

Wie bereits in der Einleitung erwähnt, gibt es wenig vergleichbare Studien und keine, die sich mit der Kombination aus Veröffentlichung und Wiederverwendung von Items be-

schäftigen. Joncas et al. (2018) konnten ebenfalls einen zunehmenden Schwierigkeitskoeffizienten bei wiederholter Verwendung von MC-Fragen nachweisen, wenn auch geringer als in dieser Studie (bis zu *d* = 0,054 bei vier Wiederverwendungen vs. bis zu *d* = 0,14 zwischen wiederverwendeten, veröffentlichten Items und der Baseline Sommersemester 2017 in dieser Studie) (10,11). Ob und wie Items hier veröffentlicht wurden, wird in der Studie und auch auf Nachfrage bei den Autoren leider nicht beschrieben. Wegen der relativen geringen Effekte und der in der Studie angegeben Bedingungen kann aber davon ausgegangen werden, dass es keine absichtliche Veröffentlichung von Items gab. Des Weiteren beobachteten Joncas et al. (2018) eine sinkende Trennschärfe mit wiederholter Verwendung der Items. Dies erklärt sich wahrscheinlich hauptsächlich aus der Berechnungsmethode: Joncas et al. (2018) benutzen den im Abschnitt 1.2.1 beschriebenen „Index of Discrimination“ während wir den korrigierten Korrelationskoeffizienten *r‘* nach Pearson verwenden, welcher allgemein weniger empfindlich für Veränderungen der Itemschwierigkeit und damit genauer ist (10).
Die anderen in der Einleitung beschriebenen Arbeiten sind nicht mit unserer zu vergleichen, da sie deutlich kleinere Kollektive an Items und Prüflingen untersuchten (27–29) oder Fragen nach Veröffentlichung nicht wiederverwendet wurden (17).

4.4 Stärken und Schwächen der Studie

Die große Stärke der Studie besteht in der einzigartigen, quasi experimentellen Studiensituation, bei der die spezielle Intervention „Veröffentlichung von MC-Fragen“ im ansonsten unveränderten Setting einer medizinischen Fakultät, die auf die Wiederverwendung von MC-Fragen angewiesen ist, ausgewertet werden kann.
Die zweite große Stärke der Studie besteht im großen untersuchten Kollektiv von Fragen und teilnehmenden Studierenden, was die statistische Validität der untersuchten Parameter deutlich erhöht. Nach unserer Kenntnis ist dies die mit großem Abstand umfangreichste Analyse von Items (>10.000) und untersuchten Examina (199 Klausuren mit >23.500 Teilnehmern) in einer medizinischen Fakultät. Andere Studien schlossen maximal 1.629 (11) oder nur < 200 Items ein (27–29), bei gleichzeitig deutlich weniger Prüflingen pro Klausur.

Eine weitere Stärke besteht in der übersichtlichen Methodik, die ausschließlich routinemäßig erhobene Prüfungsparameter statistisch auswertet und so leicht von Mitarbeitenden anderer Fakultäten interpretiert werden kann.

Eine Schwäche der Studie ist die eingeschränkte Übertragbarkeit auf andere Fakultäten, insbesondere international, wo nicht immer eine der Gleitklausel ähnliche Regelung eingesetzt wird. Aber auch die interne Organisation der medizinischen Fakultäten unterscheidet sich teilweise deutlich. Die von uns beobachteten, relativ geringen Effekte der Veröffentlichung auf Bestehensquoten und Ranking der Studierenden basieren auf einer sehr großen, zentral gepflegten Datenbank mit > 20.000 Items. Wir vermuten, dass diese Datenbank momentan auf Fakultätsniveau die größte in Deutschland ist und die von uns beobachteten Effekte bei kleinerem Pool an verfügbaren Prüfungsfragen deutlicher ausfallen könnten. An manchen Fakultäten gibt auch keine zentrale Item-Datenbank und Prüfungserstellung wie an der Charité, sodass die Prüfungen stilistisch und inhaltlich inhomogen sind und sich Ergebnisse nur begrenzt übertragen lassen.

4.5 Implikationen für Praxis und zukünftige Forschung

Eine generelle Empfehlung lässt sich aus unseren Ergebnissen nicht ableiten, da diese nur begrenzt in andere Curricula übertragbar sind und die Wertung maßgeblich vom Ziel der Prüfung abhängt.

In *summativen* Prüfungen, deren Ergebnisse große Auswirkungen auf die weitere Karriere der Prüflinge haben, z.B. Staatsexamina, können die von uns beobachteten Veränderungen durch Veröffentlichung von Items bereits die Reliabilität in nicht akzeptablem Ausmaß beeinträchtigen. Gleichzeitig ist in diesen sog. „High-Stakes“-Prüfungen Transparenz über das Zustandekommen des Ergebnisses sehr wichtig, weshalb z.B. das deutsche und das koreanische Staatsexamen die Prüfungsfragen trotzdem veröffentlichen, aber eben nicht wiederverwenden. Dieses Vorgehen verbraucht viele Ressourcen und ist entsprechend auf Fakultätsniveau nicht umsetzbar. (10)

In *formativen* Prüfungen hingegen, die in erster Linie der Wissenskontrolle und der Lernmotivation dienen, könnten die Vorteile der Veröffentlichung durch verbessertes Feedback überwiegen. Wünschenswert wäre hier statt dem reinen Zur-Verfügung-Stellen der Items ein individuelles Feedback, warum eine bestimmte Antwort richtig oder falsch ist. An der Charité befindet sich ein solches Tool aktuell in Entwicklung (37). Die negativen Effekte der Veröffentlichung ließen sich über eine solche Plattform eventuell auch abmildern, wenn nur Prüflingen nach Login die verwendeten Prüfungsfragen zur Verfügung gestellt werden. Die könnten zwar immer noch kopiert werden, aber der Aufwand, diese

einzeln zu kopieren ist deutlich höher, als einfach nur die kopierte Klausur auf eine file-sharing-Plattform zu laden. Essenziell ist auch weiterhin eine ausreichend große Item-Datenbank, damit unsere Ergebnisse übertragbar bleiben. (10)

Schlussendlich kann man sich angesichts des dazu nötigen Aufwandes fragen, ob solche formativen MC-Prüfungen sich noch relevant von den bereits verwendeten Progress Tests unterscheiden (38). Diese teilen viele der in dieser Studie für die Veröffentlichung von Fragen beschriebenen Vorteile wie verminderte Prüfungsangst und vermehrte Transparenz und bieten nochmals bessere Feedback-Optionen, da sie nicht nur direktes, strukturiertes Feedback über die zurückliegende Prüfung vermitteln, sondern auch longitudinales Feedback über den Lernfortschritt über mehrere Studienjahre. Zusätzlich sind die verwendeten Item-Datenbanken oft zentralisiert und werden von mehreren Fakultäten genutzt, was Ressourcen spart. Ein Nachteil von Progress Tests besteht in der in vielen Fakultäten freiwilligen Teilnahme und der Bedeutungslosigkeit der Ergebnisse für den Studienfortschritt, was die Lernmotivation und damit die Effekte des Test-enhanced Learning schmälert. Außerdem prüfen Progress Tests in der Regel auf dem Niveau eines Studienabgängers/-abgängerin, sodass zusätzlich wenig Motivation besteht, gerade unterrichtete Inhalte zu wiederholen. Die Kombination aus formativen, fakultätsindividuellen MC-Prüfungen und einer Aufwertung der Progress Tests (z.B. durch verpflichtende Teilnahme und Mentoringgespräche bei ausbleibendem Lernfortschritt) könnte eine sinnvolle Möglichkeit darstellen, ressourcenschonend den Anforderungen eines kompetenzbasierten Curriculums gerecht zu werden. (38–40)

Angesichts der Diversität der deutschen und internationalen Curricula wären Folgestudien unter anderen Bedingungen an anderen Fakultäten sinnvoll, um unsere Ergebnisse zu verifizieren. Nicht zuletzt wäre auch eine Wiederholung der Studie an unserer Fakultät unter Einschluss weiterer Semester wünschenswert.

5. Schlussfolgerungen

Diese Studie gibt Curriculumsentwickler*innen erstmals quantitative Werte an die Hand, die sie zur Orientierung nutzen könne, um die Effekte der Veröffentlichung und Wiederverwendung von MC-Fragen an ihrer Fakultät abzuschätzen. Ob die von uns beschriebenen negativen Effekte auf die Itemschwierigkeit oder die angenommenen Vorteile einer Itemveröffentlichung überwiegen, hängt maßgeblich vom Prüfungsziel und der Einbettung im Curriculum ab und muss jeweils abgewogen werden.

Literaturverzeichnis

1. Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Advances in Health Sciences Education [Internet]. 1996 [cited 2022 Apr 16];1(1):41–67. Available from: https://link.springer.com/article/10.1007/BF00596229
2. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach [Internet]. 2010 Aug [cited 2022 Jun 29];32(8):676–82. Available from: https://pubmed.ncbi.nlm.nih.gov/20662580/
3. Lockyer J, Carraccio C, Chan MK, Hart D, Smee S, Touchie C, et al. Core principles of assessment in competency-based medical education. Med Teach [Internet]. 2017 Jun 3 [cited 2023 Feb 8];39(6):609–16. Available from: http://dx.doi.org/10.1080/0142159X.2017.1315082
4. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019 Jan 2;53(1):76–85.
5. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011 Jun 24;33(6):478–85.
6. Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48. Med Teach. 2018 Apr 3;40(4):337–50.
7. Larsen DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. Med Educ. 2008 Oct;42(10):959–66.
8. Butler AC, Roediger HL. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Mem Cognit. 2008;36(3):604–16.
9. Epstein RM. Assessment in medical education [Internet]. Vol. 356, New England Journal of Medicine. 2007 [cited 2018 Dec 13]. p. 387–96. Available from: http://www.nejm.org/doi/10.1056/NEJMra054784
10. Appelhaus S, Werner S, Grosse P, Kämmer JE. Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools. Med Educ Online [Internet]. 2023;28(1). Available from: https://doi.org/10.1080/10872981.2022.2143298

11. Joncas SX, St-Onge C, Bourque S, Farand P. Re-using questions in classroom-based assessment : an exploratory study at the undergraduate medical education level. Perspect Med Educ. 2018;7:373–8.
12. Möltner A, Schellberg D, Jünger J. Basic quantitative analyses of medical examinations. GMS Z Med Ausbild. 2006;23(3):Doc53.
13. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33(6):447–58.
14. McCrossan P, Nicholson A, McCallion N. Minimum accepted competency examination: test item analysis. BMC Med Educ [Internet]. 2022 Dec 1 [cited 2023 Apr 2];22(1):1–7. Available from: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-022-03475-8
15. FSMB/NBME 2021. Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME). Exam Security. https://www.usmle.org/step-exams/exam-security. Accessed December 22, 2021.
16. Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. Med Educ. 2016;50(1):101–14.
17. Yang EB, Lee MA, Park YS. Effects of test item disclosure on medical licensing examination. Advances in Health Sciences Education. 2018;23(2):265–74.
18. Miller GE. The assessment of clinical skills/competence/performance. Acad Med [Internet]. 1990 [cited 2023 Mar 15];65(9 Suppl):S63–7. Available from: https://pubmed.ncbi.nlm.nih.gov/2400509/
19. Panigua MA, Swygert KA. Constructing Written Test Questions For the Basic and Clinical Sciences. 4th ed. Philadelphia, PA: National Board of Medical Examiners; 2016.
20. Krebs R. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen. Bern, Switzerland: Institute for Medical Education, University of Bern, Switzerland; 2002.
21. Tonkin AL. “Lifting the carpet” on cheating in medical school exams. BMJ (Online). 2015;351(August).
22. Park YS, Yang EB. Three controversies over item disclosure in medical licensure examinations. Med Educ Online. 2015;20:1, 28821.
23. Quek TTC, Tam WWS, Tran BX, Zhang M, Zhang Z, Ho CSH, et al. The global prevalence of anxiety among medical students: A meta-analysis. Vol. 16, International Journal of Environmental Research and Public Health. MDPI AG; 2019. p. 2735.

24. Wadi M, Yusoff MSB, Abdul Rahim AF, Lah NAZN. Factors affecting test anxiety: a qualitative analysis of medical students' views. BMC Psychol. 2022 Dec 1;10(1):8.
25. Guraya SY, Guraya SS, Habib F, AlQuiliti KW, Khoshhal KI. Medical students' perception of test anxiety triggered by different assessment modalities. Med Teach. 2018 Jul 6;40(sup1):S49–55.
26. Gilmer JS. The Effects of Test Disclosure on Equated Scores and Pass Rates. 1983;245–55.
27. Wagner-Menghin M, Preusche I, Schmidts M. The Effects of Reusing Written Test Items : A Study Using the Rasch Model. ISRN Education. 2013;2013:Article ID 585420.
28. Herskovic P. Reutilization of multiple-choice questions. Med Teach. 1999 Jan 3;21(4):430–1.
29. Wood TJ. The effect of reused questions on repeat examinees. Advances in Health Sciences Education. 2009;14:465–73.
30. Stricker LJ. Test Disclosure and Retest Performance on the SAT. Appl Psychol Meas. 1984;8(1):81–7.
31. Bestehens- und Notengrenzen. Institut für medizinische und pharmazeutische Prüfungsfragen, Mainz. [Accessed December 16, 2020]. https://www.impp.de/pruefungen/allgemein/bestehens-und-notengrenzen.html.
32. Möltner A. Dealing with flawed items in examinations: Using the compensation of disadvantage as used in German state examinations in items with partial credit scoring. GMS J Med Educ. 2018;35(4):Doc49.
33. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge Academic; 1988.
34. diagrammeditor.de [Web Page / Computer software]. Vienen, NL: Zygomatic.
35. SPSS Statistics for Windows[Computer software]. Version 25.0. Armonk, NY: IBM Corp; 2017. IBM Corp., Armonk, NY, USA;
36. GraphPad Prism [Computer software]. Version 9.1. San Diego, CA: GraphPad Software; 2021.
37. Roa Romero Y, Tame H, Holzhausen Y, Petzold M, Wyszynski JV, Peters H, et al. Design and usability testing of an in-house developed performance feedback tool for medical students. BMC Med Educ. 2021;21(1):1–9.

38. Nouns ZM, Georg W. Progress testing in German speaking countries. Med Teach. 2010;32(6):467–70.
39. van der Vleuten C, Freeman A, Collares CF. Progress test utopia. Perspect Med Educ. 2018;7(2):136–8.
40. Pugh D, Regehr G. Taking the sting out of assessment: is there a role for progress testing? Med Educ. 2016;50(7):721–9.

Eidesstattliche Versicherung

„Ich, Stefan Appelhaus, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: *Wiederverwendung und Veröffentlichung von Multiple-Choice-Fragen in medizinischen Prüfungen – Auswirkungen auf psychometrische Kennwerte und Klausurergebnisse / Reuse and Disclosure of Multiple Choice Questions in Medical School Exams – Effects on Item Psychometrics and Test results* selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit der Erstbetreuerin, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.org) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.
Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Anteilserklärung an den erfolgten Publikationen

Stefan Appelhaus hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1:
Stefan Appelhaus, Susanne Werner, Pascal Grosse, Juliane E. Kämmer. *Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools*. Medical Education Online, 28:1. Veröffentlicht: 09. November 2022.
Beitrag im Einzelnen:
Während seiner Arbeit im Prüfungsbereich der Charité begann sich Stefan Appelhaus für die Frage zu interessieren, welche Auswirkungen die Entscheidung zur Quasi-Veröffentlichung der Prüfungsfragen auf die psychometrischen Parameter der einzelnen Items haben könnte. Als er im Rahmen einer Literaturrecherche keine zufriedenstellende Antwort fand, entwickelte er die Idee für diese Dissertation, bemühte sich um eine geeignete Betreuung der Promotion, entwickelte die hier und in der zugehörigen Publikation bearbeiteten Fragestellungen und Methodiken, sammelte die notwendigen Daten, bereitete sie auf und wertete sie statistisch aus. Er entwarf das Manuskript, erstellte die Abbildungen und Tabellen und reichte die finale Version beim Verlag ein. Die Koautoren und Prof. Dr. Liane Schenk als Erstbetreuerin der Promotion standen ihm während des gesamten Prozesses unterstützend zur Seite.

Publikation 2:
Stefan Appelhaus, Juliane E. Kämmer, Susanne Werner. *Auswirkungen auf das Antwortverhalten bei Wiederverwendung von MC-Prüfungsfragen*. Jahrestagung der Gesellschaft für medizinische Ausbildung (GMA). Zürich, Schweiz. 15. September 2021.
Beitrag im Einzelnen:
Erarbeitung von Fragestellung und Methodik sowie Erarbeitung und Auswertung des Datensatzes wie unter Publikation 1 beschrieben. Zusammenstellung der Präsentationsfolien. Die Vorstellung der Ergebnisse erfolgte aus persönlichen Gründen vertretungsweise durch Susanne Werner.

_____

Unterschrift, Datum und Stempel der erstbetreuenden Hochschullehrerin

_____

Unterschrift des Doktoranden

Druckexemplar der Publikation

MEDICAL EDUCATION ONLINE
2023, VOL. 28, 2143298
https://doi.org/10.1080/10872981.2022.2143298

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools

Stefan Appelhaus [a,b], Susanne Werner [c], Pascal Grosse [d] and Juliane E. Kämmer [e]

[a]Institute of Medical Sociology and Rehabilitation Science, Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany; [b]Department of Radiology and Nuclear Medicine, Universitätsmedizin Mannheim, Heidelberg University, Mannheim, Germany; [c]Assessment Unit, Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany; [d]Dean of Students Office and Department of Neurology, Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany; [e]Department of Emergency Medicine, University of Bern, Bern, Switzerland

ABSTRACT

Background: Disclosure of items used in multiple-choice-question (MCQ) exams may decrease student anxiety and improve transparency, feedback, and test-enhanced learning but potentially compromises the reliability and fairness of exams if items are eventually reused. Evidence regarding whether disclosure and reuse of test items change item psychometrics is scarce and inconclusive.

Methods: We retrospectively analysed difficulty and discrimination coefficients of 10,148 MCQ items used between fall 2017 and fall 2019 in a large European medical school in which items were disclosed from fall 2017 onwards. We categorised items as 'new'; 'reused, not disclosed'; or 'reused, disclosed'. For reused items, we calculated the difference from their first ever use, that is, when they were new. Differences between categories and terms were analysed with one-way analyses of variance and independent-samples *t* tests.

Results: The proportion of reused, disclosed items grew from 0% to 48.4%; mean difficulty coefficients increased from 0.70 to 0.76; that is, items became easier, $P < .001$, $\eta_p^2 = 0.011$. On average, reused, disclosed items were significantly easier ($M = 0.83$) than reused, not disclosed items ($M = 0.71$) and entirely new items ($M = 0.66$), $P < .001$, $\eta_p^2 = 0.087$. Mean discrimination coefficients increased from 0.21 to 0.23; that is, item became slightly more discriminating, $P = .002$, $\eta_p^2 = 0.002$.

Conclusions: Disclosing test items provides the opportunity to enhance feedback and transparency in MCQ exams but potentially at the expense of decreased item reliability. Discrimination was positively affected. Our study may help weigh advantages and disadvantages of using previously disclosed items.

ARTICLE HISTORY
Received 26 August 2022
Revised 30 October 2022
Accepted 31 October 2022

KEYWORDS
Assessment; formative assessment; multiple-choice questions; item reuse; disclosure; feedback; progress testing

Introduction

With the continuing trend towards competency-based curricula and formative assessment in medical education, the expectations for assessment have been growing. Rather than just reliably measuring students' performance, assessment also has to provide feedback about the learners' strengths and weaknesses, enhance learning, and steer their learning process, often summarised by the term test-enhanced learning [1–3]. To fulfil these expectations, many new assessment formats such as objective structured clinical examinations and short-answer questions have been developed [4]. Unfortunately, the development and realisation of these new assessment formats require many resources. Thus, written examinations using multiple-choice questions (MCQs) are still a key element in the assessment of medical students across the world [5]. To use MCQs in frequent formative assessments, items are needed in great numbers. To achieve this goal many medical schools around the world reuse items, as resources and personnel are usually limited [5–7]. To ensure that the psychometric properties of MCQs are maintained, examiners go to great lengths to keep items confidential [8]. This strict confidentiality is especially important for summative assessments, such as in high-stakes medical licensure examinations (e.g., the USA Medical Licensure Examination), but it does not allow for providing feedback to examinees. Feedback is, however, one of the crucial properties of formative assessment, as it can foster test-enhanced learning and channel long-term retention of course content [9,10]. Thus, the possibility of providing feedback is one argument for item disclosure.

CONTACT Stefan Appelhaus stefan.appelhaus@charite.de Institute of Medical Sociology and Rehabilitation Science, Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10872981.2022.2143298

Further arguments for disclosure include the increased overall transparency of the process of administering exams, the reduced effort required of examiners to ensure test items are kept confidential, the opportunity for examinees to better prepare for the test, and the reduction of student anxiety [11–14]. Conversely, the main argument against disclosure is that examinees may more easily obtain previous original test items so that they will eventually test better than earlier cohorts and, in consequence, compromise reliability [7].

To date, evidence for the latter argument has been scarce and contradictory [11]. For example, Yang et al. investigated effects of disclosure of items, answers, and performance data in South Korea's medical licensing examination [15]. They found no significant changes in student performance, pass rates, or item psychometrics but they did not reuse previously disclosed items. Herskovic disclosed items by discussing them after the exam with examinees without giving them copies of the items to keep [16]. He found no relevant change in psychometric parameters when items were reused in following exams. Wood stated that repeat candidates did not have better scores for those items they had answered before compared to unknown items [17]. However, the sample size in both studies (Herskovic: 197 reused items, number of examinees not stated; Wood: 26 items, 130 examinees) was rather small. Joncas et al. exclusively studied the changes in psychometric parameters over a period of five years during which 1,629 items were reused up to four times in a Canadian medical school without official disclosure of items. They showed that each reuse of items led to a decline in difficulty and discrimination [6]. This deterioration of item psychometrics supports the assumption that it is common for examinees in many medical schools to try to obtain exam questions from previous examinees [7]. It remains unclear whether official disclosure of items would have an additional effect on item psychometrics if students apparently have access to original items anyway.

Hence, given the widespread use of MCQ items and the inconsistent evidence concerning the effects of disclosure and reuse, a more systematic analysis of the problem is obviously a desideratum. Our study addresses this concern, making use of a quasi-experimental setting in one of the largest medical faculties in Europe. For reasons of practicality and transparency, our medical school implemented a radical policy change regarding the disclosure of items. Changing from a previously very restrictive practice that allowed no disclosure of items whatsoever, from fall 2017 students were allowed to take home their tests and were provided with the answer keys, thus making it much easier to provide feedback. Obviously, subsequent students faced no challenge obtaining the original items. Before fall 2017, students had to turn in their tests so that they could, at best, only try to memorise the items, and answer keys were not published. Because of this policy change, the percentage of reused, disclosed items gradually grew from 0% to 48% per exam between fall 2017 and fall 2019.

This policy change offers an exceptional opportunity to examine the effect of disclosure and reuse of test items by comparing the results of exams prior to and after the disclosure of test items. Such a constellation comes close to an experimental setting in which the results prior to disclosure act as the control group. Using this methodological approach, we analysed potential differences in item psychometrics between (a) reused, disclosed items, (b) reused, not disclosed items, and (c) entirely new items. To better assess the effect of disclosure, we analysed the changes in item psychometrics of reused (both, disclosed and not disclosed) items compared to their first ever use (i.e., when they were new items). In line with previous, smaller scale research [6], we hypothesised that items would become easier to solve correctly and less discriminating on reuse and even more so on reuse after disclosure, as students theoretically had no access to original test items before this policy change and would now possibly obtain them more easily.

Methods

Study setting

The study was conducted at *Charité – Universitätsmedizin Berlin*, one of Europe's largest medical schools with approximately 4,500 medical students. At the end of each term (two terms per year), all students from their first to their fifth year take an end-of-term exam that uses exclusively one-best-answer MCQs. As the curriculum is organised in a modular fashion, each exam is interdisciplinary, covering basic sciences and clinical disciplines. For the time our study covers, each exam comprised between 40 and 120 items, depending on the term. Each exam was offered twice per term; items used in the first exam could not be used in the second. All exams were paper based. Prior to 2017, exam copies were collected after the exam and answer keys were not published. Since 2017, students have been allowed to take their individual copies home and answer keys have been published online.

All exams are developed according to a standardised procedure: The learning aims are selected randomly but are nevertheless based on a blueprint so that the selection of topics tested in the exam is representative of the whole module. For up to 80% of these selected learning aims, test items that have already been used are randomly selected from a database containing more than 20,000 items

developed by faculty over almost a decade. The items for the remaining 20% of the learning aims are developed entirely anew by faculty members. To ensure high-quality exams, items are written only by experts in their respective fields and new items and final exams are proofread by at least two faculty members and one specialist responsible for linguistic and formal quality control of test items. In a post-examination review, students can report potentially flawed questions and the Board of Examiners may change scoring.

Exams at *Charité* are graded according to the principles of the German national licensing exams. Exam grades range between 1 and 5, with 1 being the best and 4 indicating the lowest passing grade. With a grade of 5, students fail the exam. The passing score is 60%. The exam can be considered high stakes as students are not allowed to further pursue their studies without passing the exam; yet, they have up to six tries to do so.

Study methods

We included all exams used in the five terms between fall 2017 and fall 2019. Note that we did not include the terms after 2019 because assessment practices changed in the course of the COVID-19 pandemic starting in 2020. The exams in fall 2017 were the first to be disclosed. As these exams did not include reused, disclosed items, we used this examination period as a baseline. Only the four examination periods from spring 2018 to fall 2019 included previously disclosed items, naturally in growing proportions.

We assessed the number of participants as well as item difficulty and discrimination coefficients. Item difficulty was determined as the number of students who answered the question correctly divided by the number of all participating students. Results for difficulty range between 0 and 1 with higher values indicating easier items. According to classic test theory, item difficulty coefficients should ideally range from 0.4 to 0.8 [18]. Discrimination was estimated as a Pearson correlation coefficient of students' performance on the item with their overall score in the exam [18]. Results for discrimination range between −1 and 1 with higher values representing a higher degree of discrimination. Möltner et al. considered a discrimination coefficient of 0.2 or higher acceptable for MCQ items [18]. For difficulty coefficients of 1, it is not possible to calculate discrimination coefficients.

We grouped items according to the date of their last use in an exam into one of the following three categories:

New: Never used before this exam.

Reused, not disclosed: Last used before disclosure of exam content (i.e., last used in spring 2017 or earlier).

Reused, disclosed: last used after disclosure of exam contents (i.e., last used in fall 2017 or later).

We excluded items if the post-examination review deemed them to be inadequate. Note that items may have been used more than once in the study period; in this case they were included for each use in the respective category.

Analysis

To assess changes in overall mean item psychometrics, we ran two one-way analyses of variance (ANOVAs) with term (fall 2017, spring 2018, fall 2018, spring 2019, and fall 2019) as independent variable and difficulty and discrimination coefficients as dependent variables. To assess differences in item psychometrics between item groups, we ran two one-way ANOVAs with item group as independent variable and difficulty and discrimination coefficients as dependent variables.

To assess the relative impact of reuse without disclosure versus reuse after disclosure on item difficulty and discrimination, we calculated the differences in difficulty and discrimination coefficients (a) between an item's first use (i.e., when it was a new item) and when it was reused (delta reused, not disclosed – new), and (b) between an item's first use and when it was reused and disclosed (delta reused, disclosed – new). We entered these differences into an independent-samples t test.

Whenever ANOVA results were significant, Bonferroni post hoc tests were used to assess differences between groups. Effect sizes for ANOVAs were calculated as η_p^2. We follow Cohen's [19] definition of benchmarks for small ($\eta_p^2 = 0.01$), medium ($\eta_p^2 = 0.06$), and large ($\eta_p^2 = 0.14$) effects. Effect sizes for t tests were calculated as Cohen's d with defined benchmarks for small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effects. All calculations were done using IBM SPSS 25 [20] and figures were created using GraphPad Prism 9 [21].

Results

We included the results of 199 exams with 23,507 participants and 10,148 items in our analyses (see Figure 1 for an overview of items). Percentage of repeat examinees in the study period was 4.8%.

Although the percentage of reused, disclosed items steadily increased up to 48.4% in fall 2019 (Figure 2a), item psychometrics varied only to a small degree between terms: Mean difficulty coefficients ranged from 0.70 in fall 2017 to 0.76 in spring 2019, $F(4, 10{,}143) = 27.03$, $P < .001$, $\eta_p^2 = 0.011$ (Figure 2b). Mean discrimination coefficients ranged from 0.21 in fall 2017 to 0.23 in spring 2019, $F(4, 9{,}734) = 4.25$,

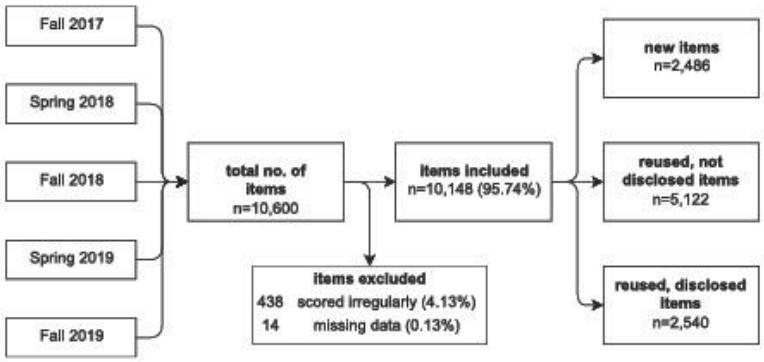

Figure 1. **Overview of origin and types of items.** We analysed all exams conducted between the decision to disclose multiple-choice questions in future examinations in fall 2017 and fall 2019. Fall 2017 was the first exam disclosed to examinees but without reuse of previously disclosed items, thus constituting our baseline for comparison with the exams from spring 2018 to fall 2019, which included reused, previously disclosed items in growing proportions.

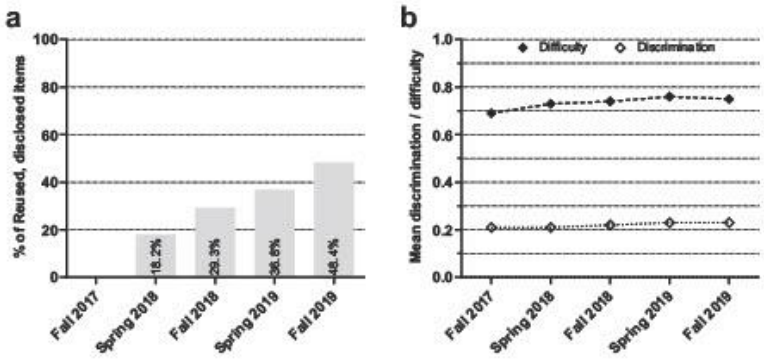

Figure 2. **Item psychometrics over terms.** (a) Proportion of reused, disclosed items across terms. Note that fall 2017 served as baseline without any disclosed items. (b) Mean difficulty and discrimination coefficients, across all item groups, per term.

$P = .002$, $\eta_p^2 = 0.002$ (Figure 2b). Including item groups as an additional independent variable in the analyses revealed that only reused, not disclosed items slightly changed in difficulty over terms. A detailed overview of this additional analysis and all other results can be found in Supplement Table 1.

The ANOVA revealed differences in difficulty of medium effect size between item groups, $F(2, 10{,}145) = 483.38$, $P < .001$, $\eta_p^2 = 0.087$. New items were most difficult ($M = 0.66$) followed by reused, not disclosed items ($M = 0.71$) and reused, disclosed items ($M = 0.83$). Bonferroni post hoc tests revealed differences between all groups (all $P < .001$). Mean discrimination coefficients varied between item groups, $F(2, 9{,}736) = 49.50$, $P < .001$, $\eta_p^2 = 0.008$, though the effect size was negligible. New items ($M = 0.20$) and reused, not disclosed items ($M = 0.21$) were less discriminating than reused, disclosed items ($M = 0.25$). Bonferroni post hoc tests revealed differences only between reused, disclosed items and the other two item groups (both $P < .001$). Results are presented in Figure 3.

To better assess the effect of disclosure on items, we focused on the groups of reused items and analysed their item psychometrics as compared to their first ever use, that is, when they were new items. Difficulty coefficients decreased by $M = -0.01$ in reused, not disclosed items and increased by $M = 0.11$ in reused, disclosed items, indicating a medium to large effect of disclosure on item difficulty, $t(4699.79) = -29.69$, $P < .001$, $d = 0.74$. Discrimination coefficients increased by $M = 0.03$ in reused, not disclosed items and by $M = 0.07$ in reused, disclosed items, indicating a negligible to small effect of disclosure on item discrimination, $t(4274.07) = -7.22$, $P < .001$, $d = 0.19$.

We did not include students grades and pass rates in our analysis because our medical school applies an automatic adjustment clause used in the German state examinations [22,23]. The latter procedure is meant to level grades across terms and to prevent

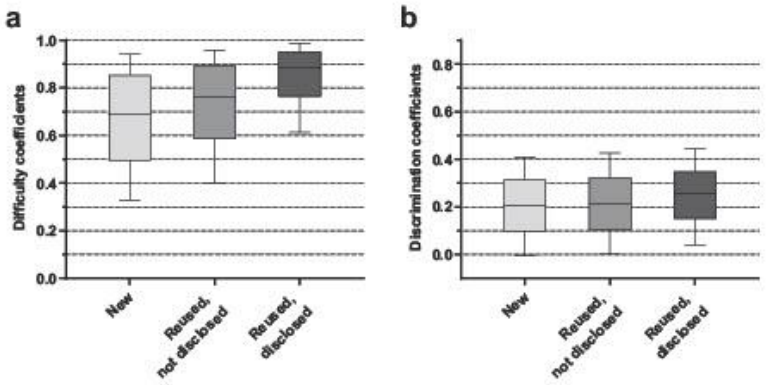

Figure 3. Boxplot diagram of difficulty and discrimination coefficients between item groups. Boxplot whiskers mark the 10th and 90th percentiles of the data. (a) Average difficulty coefficients per item group, across terms (including baseline). (b) Average discrimination coefficients per item group, across terms (including baseline).

unusually high failure rates in the high-stakes German medical licensure examination. Mean grades and pass rates are listed in Supplement Table 1. As expected, they are barely affected.

Discussion

Implementing test-enhanced learning strategies in a curriculum often requires many resources, which may deter medical schools from doing so [1]. By simply disclosing existing items and exams to students, examiners can provide feedback with relatively low effort. Furthermore, withholding exam items from students may lead them to doubt exam transparency and fairness. Because of the need for efficient and frequent assessments under conditions of limited resources, many medical schools rely on item reuse [6,7]. However, the combination of item disclosure and reuse facilitates cheating in the form of content sharing with subsequent examinees, thus potentially jeopardising item validity and exam reliability [7]. Before considering providing feedback through item disclosure, examiners need to know the effects of this measure on item psychometrics. Yet, empirical evidence on the consequences of disclosing exam items has been scarce and mostly contradictory, which reflects the obvious methodological difficulties encountered in addressing such a research question [6,11,15–17]. Our almost unique and quasi-experimental setting, which came about more or less by chance, nevertheless allowed us to investigate the effects of reusing non-disclosed and disclosed test items on item difficulty and discrimination in a large European medical school in a high-income country with a large number of participants and including more than 10,000 test items.

As we hypothesised, our analyses revealed that reused, disclosed items were easier to answer correctly than reused, not disclosed and new items. Also, reused, disclosed items, but not reused, not disclosed items, were answered more accurately (+0.11 in the difficulty coefficient) than when they were used for the first time. These results suggest, first, that the observed change in difficulty coefficients is indeed mostly due to disclosure of items and students using the disclosed items to prepare for their exams. Second, previous measures to keep items confidential seem to have been mostly effective.

Our analyses further revealed, counterintuitively, that reused, disclosed items were slightly more discriminating than reused, not disclosed and new items and also compared to when they were used for the first time. Our finding is in contrast to the results of a recent study [6] that showed a decrease in discrimination with each reuse of an item. The discrepancy might be due to different methods of calculating discrimination coefficients. Whereas Joncas et al. used the item-discrimination index method, we used the point-biserial correlation coefficient, which is generally less affected by decreasing item difficulty [18,24]. Another explanation would be that higher performing students can better memorise a higher number of items, leading to increased discrimination. We conclude that items still discriminate correctly between higher and lower performing students, and, thus, item disclosure seems not to interfere with correct student ranking.

We suspect that the size of the effects item disclosure and reuse have on difficulty and discrimination depends on the size of the item pool from which exam items are drawn. Our medical school's item bank contains more than 20,000 MCQs altogether, which translates to roughly 2,000 potential items per exam, of which only a few are randomly selected for the end-of-term exam. It seems obvious that it is

almost impossible for students to perfectly memorise this many items. Naturally, the smaller an item bank is, the easier it is for students to memorise items and the faster the proportion of disclosed items increases. Thus, the effects of item disclosure may be more pronounced in institutions that do not generate as many test items as our institution has done. To oppose this effect, institutions could attempt to enlarge their item banks. To do so, medical schools have tried innovative approaches such as developing items using artificial intelligence [25] or having students write items as part of the coursework. Especially the latter has been shown to have positive effects on learning, too [26,27].

In sum, our results show that increased transparency and feedback in MCQ exams come at the expense of item reliability. Whether the size of this effect is practically relevant most likely depends on the goals of exams and their underlying philosophy [2]. In high-stakes, career-deciding exams such as medical licensure examinations, even small changes in reliability induced by disclosure and reuse of items may be too big to tolerate. In formative assessment settings in which discrimination and feedback are more important than weeding out students through reliable passing scores, a drop in item psychometrics may be outweighed by the benefits of disclosure and reuse.

Disclosing items for the purpose of providing feedback can take many forms, from simply publishing exams and answer keys to incorporating individualised feedback and explanations on systematic feedback platforms (e.g., [28],). Yet, developing such feedback platforms requires many resources, which may be especially problematic for smaller medical schools. Thus, progress testing may be a viable alternative to frequent assessment and disclosing items [29,30], as progress testing shares many of the advantages of item disclosure, such as better feedback options, increased transparency, and reduced student anxiety, but also the ensuing need for large item banks. Progress testing has the additional advantage of providing systematic individual and longitudinal feedback, which likely entails much richer information than just providing items and answer keys without explanation [31,32].

It is an empirical question whether our results can be generalised to other medical schools with perhaps smaller item banks, different examination schedules, alternative blueprints for their exams, or different methods to prevent leakage of items. Nevertheless, our study could potentially provide a more general lesson regarding the development, use, and reuse of MCQs. It would be rewarding if our study could inspire further studies in various academic settings and in countries with other academic traditions. Finally, it would be of particular interest to investigate whether the effects delineated in our study can be classified as short-term or sustainable effects. To answer the latter question, follow-up studies at our institution covering a longer time span are mandatory. As the assumed positive effects of disclosure, for example, reduced anxiety and test-enhanced learning, have been shown only in other settings, student surveys or correlation with dropout rates and licensure examination results would be interesting for further research.

Conclusion

Our study provides evidence supporting the argument that item disclosure in combination with item reuse decreases item difficulty and increases discrimination. Thus, disclosure may compromise exam reliability to a moderate extent. Our results may help educators weigh the observed disadvantages of item disclosure against the obvious benefits such as increased transparency, reduced student anxiety, and the opportunity to provide better feedback.

Acknowledgments

We would like to thank all current and former faculty involved with the generation of MCQs and administration of exams at Charité—Universitätsmedizin Berlin for participation in the acquisition and provision of data. We are grateful to Anita Todd for language editing the manuscript. We acknowledge financial support from the Open Access Publication Fund of Charité—Universitätsmedizin Berlin and the German Research Foundation (DFG).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethical approval

This study was approved by the ethics committee of *Charité—Universitätsmedizin Berlin*. (EA4/198/20)

ORCID

Stefan Appelhaus http://orcid.org/0000-0002-1325-6422
Susanne Werner http://orcid.org/0000-0002-7710-6677
Pascal Grosse http://orcid.org/0000-0002-0114-0430
Juliane E. Kämmer http://orcid.org/0000-0001-6042-8453

References

[1] Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: a systematic review: BEME Guide No. 48. Med Teach. 2018;40(4):337–350.
[2] Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019;53(1):76–85.
[3] Boulet JR, Durning SJ. What we measure … and what we should measure in medical education. Med Educ. 2019;53(1):86–94.
[4] Sam AH, Wilson R, Westacott R, et al. Thinking differently–Students' cognitive processes when answering two

different formats of written question. Med Teach. 2021;43(11):1278–1285.
[5] Epstein RM, Cox M, Irby DM. Assessment in medical education [Internet]. N Engl J Med. 2007;356:387–396.
[6] Joncas SX, St-Onge C, Bourque S, et al. Re-using questions in classroom-based assessment: an exploratory study at the undergraduate medical education level. Perspect Med Educ. 2018;7(6):373–378.
[7] Tonkin AL. "Lifting the carpet" on cheating in medical school exams. BMJ. 2015 ;351:h4014.
[8] FSMB/NBME. Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME). Exam Security. 2021. [cited 2021 Dec 22]. Available from: https://www.usmle.org/step-exams/exam-security.
[9] Larsen DP, Butler AC, Roediger IIIHL. Test-enhanced learning in medical education. Med Educ. 2008;42(10):959–966.
[10] Butler AC, Roediger HL. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Mem Cogn. 2008;36(3):604–616.
[11] Park YS, Yang EB. Three controversies over item disclosure in medical licensure examinations. Med Educ Online. 2015;20(1):28821.
[12] Wadi M, Yusoff MSB, Abdul Rahim AF, et al. Factors affecting test anxiety: a qualitative analysis of medical students' views. BMC Psychol. 2022;10(1):8.
[13] Guraya SY, Guraya SS, Habib F, et al. Medical students' perception of test anxiety triggered by different assessment modalities. Med Teach. 2018;40(sup1):S49–S55.
[14] Encandela J, Gibson C, Angoff N, et al. Characteristics of test anxiety among medical students and congruence of strategies to address it. Med Educ Online. 2014;19(1):25211.
[15] Yang EB, Lee MA, Park YS. Effects of test item disclosure on medical licensing examination. Adv Heal Sci Educ. 2018;23(2):265–274.
[16] Herskovic P. Reutilization of multiple-choice questions. Med Teach. 1999;21(4):430–431.
[17] Wood TJ. The effect of reused questions on repeat examinees. Adv Heal Sci Educ. 2009;14(4):465–473.
[18] Möltner A, Schellberg D, Jünger J. Basic quantitative analyses of medical examinations. GMS Z Med Ausbild. 2006;23:Doc53.
[19] Cohen J. Statistical power analysis for the behavioral sciences. New York: NY: Routledge Academic; 1988.
[20] SPSS Statistics for Windows [Computer software]. Version 25.0. Armonk: NY: IBM Corp; 2017.
[21] GraphPad Prism [Computer software]. Version 9.1. San Diego. CA: GraphPad Software; 2021.
[22] Bestehens- und Notengrenzen. Institut für medizinische und pharmazeutische Prüfungsfragen, Mainz. [cited 2020 Dec 16]. Available from: https://www.impp.de/pruefungen/allgemein/bestehens-und-notengrenzen.html.
[23] Möltner A. Dealing with flawed items in examinations: using the compensation of disadvantage as used in German state examinations in items with partial credit scoring. GMS J Med Educ. 2018;35(4):Doc49.
[24] Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33(6):447–458.
[25] Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. Med Educ. 2012;46(8):757–765.
[26] Herrero JI, Lucena F, Quiroga J. Randomized study showing the benefit of medical study writing multiple choice questions on their learning. BMC Med Educ. 2019;19(1):42.
[27] Touissi Y, Hjiej G, Hajjioui A, et al. Does developing multiple-choice questions improve medical students' learning? A systematic review. Med Educ Online. 2022;27:1.
[28] Roa Romero Y, Tame H, Holzhausen Y, et al. Design and usability testing of an in-house developed performance feedback tool for medical students. BMC Med Educ. 2021;21(1):1–9.
[29] van der Vleuten C, Freeman A, Collares CF. Progress test utopia. Perspect Med Educ. 2018;7(2):136–138.
[30] Pugh D, Regehr G. Taking the sting out of assessment: is there a role for progress testing? Med Educ. 2016;50(7):721–729.
[31] Nouns ZM, Georg W. Progress testing in German speaking countries. Med Teach. 2010;32(6):467–470.
[32] Kämmer JE, Hautz WE, März M. Self-monitoring accuracy does not increase throughout undergraduate medical education. Med Educ. 2020;54(4):320–327.

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Komplette Publikationsliste

Journalartikel:

Appelhaus S, Werner S, Grosse P, Kämmer JE. Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools. Med Educ Online. 2023;28(1). Available from: https://doi.org/10.1080/10872981.2022.2143298

Impact Factor: 6,0

Kongressbeiträge:

Stefan Appelhaus, Juliane E. Kämmer, Susanne Werner. *Auswirkungen auf das Antwortverhalten bei Wiederverwendung von MC-Prüfungsfragen*. Jahrestagung der Gesellschaft für medizinische Ausbildung (GMA). Zürich, Schweiz. 15. September 2021.

Danksagung

Ich bedanke mich bei meinen Betreuern Prof. Dr. Liane Schenk, PD Dr. Pascal Grosse und Dr. Juliane Kämmer sowie Susanne Werner für die hervorragende Unterstützung während der Promotion. Ohne ihre Hilfe wäre diese Arbeit in der vorliegenden Form nicht möglich gewesen.

Ich bedanke mich bei allen aktuellen und ehemaligen Mitarbeitenden des Prüfungsbereichs der Charité für Ihren Beitrag bei der Durchführung der vergangenen Prüfungen und die Möglichkeit zur Datenauswertung.

Ich bedanke mich bei meinen Eltern, Sabine und Paul, für die jahrelange, vielfältige Unterstützung während meines Studiums und darüber hinaus.

Ich bedanke mich bei Sina, Zora, Ronja und Alice für den Rückhalt und das Verständnis, mich auch in privat und beruflich ereignisreichen Zeiten dieser Dissertation widmen zu können.