Aus der Klinik für Kardiologie und Pneumologie
(Prof. Dr. med. G. Hasenfuß)
der Medizinischen Fakultät der Universität Göttingen

Vergleich der SBA- und VSA-Frageformate in der zahnmedizinischen Lehre hinsichtlich des Schweregrades, der Trennschärfe und deren Akzeptanz

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
für Zahnmedizin
der Medizinischen Fakultät der
Georg-August-Universität zu Göttingen

vorgelegt von

Franziska Jasnoch

aus

Münster

Göttingen 2021

Dekan:	Prof. Dr. med. W. Brück

Betreuungsausschuss

Betreuer/in:	Prof. Dr. med. T. Raupach
Ko-Betreuer/in:	Prof. Dr. med. dent. R. Bürgers

Prüfungskommission

Referent/in:	Prof. Dr. med T. Raupach
Ko-Referent/in:	Prof. Dr. med. dent. R. Bürgers
Drittreferent/in:	Prof. Dr. med. R. Dressel

Datum der mündlichen Prüfung:	06.09.2022

Hiermit erkläre ich, die Dissertation mit dem Titel "Vergleich der SBA- und VSA-Frageformate in der zahnmedizinischen Lehre hinsichtlich des Schweregrades, der Trennschärfe und deren Akzeptanz" eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Göttingen, den …… ……

(Unterschrift)

In dieser Arbeit wird die männliche oder neutrale Form personenbezogener Substantive gewählt. Dies impliziert keine Benachteiligung eines Geschlechts.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

α	Cronbach's Alpha
DIPS	Digitales Prüfungs- und Schulungszentrum
IMS	*item management system*
MW	Mittelwert
R	Reliabilität
SBA	*single-best-answer* (Einfachauswahl-Antworten)
SD	Standardabweichung
UMG	Universitätsmedizin Göttingen
UCAN	Umbrella Consortium for Assessment Networks
VSA	*very-short-answer* (sehr kurze Freitext-Antworten)
WS	Wintersemester

1 Einleitung

Die Evaluation der Studierenden ist ein fundamentaler Aspekt der medizinischen Lehre und muss hohen Ansprüchen genügen. Durch ihre korrekte Ausführung kann die Motivation der Studierenden zum Eigenstudium gefördert werden. Die Auswertung der Prüfungsergebnisse dient dem Lehrpersonal dabei als unmittelbares Feedback (Haigh 2003). Bereits in der Präambel der Prüfungsrichtlinie der Universitätsmedizin Göttingen (UMG) heißt es entsprechend als verbindliche Weisung:

> „Prüfungen beeinflussen das Lernverhalten positiv, wenn sie fair, zuverlässig und valide sind, korrekt durchgeführt und objektiv bewertet werden. Außerdem müssen sie die angestrebten Lernziele kongruent abbilden. Bleiben Prüfungen hinter diesen Anforderungen zurück, werden unter Umständen nur unzureichende Kenntnisse und Fertigkeiten erworben und letztlich die Lehre im betroffenen Fach oder Querschnitt entwertet." (Prüfungsrichtlinie UMG, Präambel 2020)

Deshalb ist eine genaue Analyse der Prüfungsergebnisse und der angewandten Prüfungsformate zur Qualitätssicherung der medizinischen und zahnmedizinischen Lehre erforderlich (van Bruggen et al. 2012).

Zum Testen von Wissen werden in der zahnmedizinischen Lehre fast ausschließlich *single-best-answer* (SBA)-Items verwendet. Da das SBA-Format jedoch der Kritik unterliegt, korrekte Antworten für den Prüfling trotz Unwissens mittels Ausschlussverfahren oder Hinweisen erkennbar zu machen (Elstein 1993), möchten wir – angelehnt an die Vorstudien von Sam et al. (2016, 2018, 2019a) – in der vorliegenden Studie das Prüfungswissen der Zahnmedizinstudierenden durch Integration eines *very-short-answer* (VSA)-Formates untersuchen. Außerdem soll das VSA-Format anhand grundlegender quantitativer Analysen genauer erforscht werden.

Im folgenden Abschnitt wird zunächst der aktuelle Forschungsstand in Bezug auf die allgemeine Bedeutung von Prüfungsformaten und deren Auswirkung auf das Lernverhalten der Studierenden in der medizinischen Lehre dargestellt. Im Weiteren wird ein Überblick über vorangegangene, dieser Forschungsarbeit als Grundlage dienende Studien vermittelt. Darüber hinaus werden wichtige Bestandteile der Prüfungsqualitäten genauer erläutert.

1.1 Assessment drives learning

Medizinische Kompetenz ist keine unmittelbar erbrachte Leistung, die durch rein didaktische Mittel in der medizinischen Lehre erlernt werden kann. Neben umfangreichen praktischen Erfahrungen und fortlaufenden Schulungen haben vorherige Lebenserfahrungen einen ebenso großen Einfluss auf den Kompetenzerwerb (Leach 2002). Medizinisches Personal

benötigt zur Ausübung der Profession ein umfangreiches inhaltliches Wissen, Kommunikationsfähigkeiten, ein hohes Maß an Professionalität sowie zwischenmenschliche Fähigkeiten zur Teamarbeit und im Umgang mit Patienten (Batalden et al. 2002). Eine klare Definition medizinischer Kompetenz lässt sich aufgrund ihrer Komplexität nur schwer formulieren. Entsprechend gestaltet sich auch die Testierung dieser vielfältigen Fähigkeiten generell schwierig und es bedarf einiger Veränderungen der Prüfungsmethoden.

In medizinischer und zahnmedizinischer Lehr- und Lernforschung wird den auf Lehrinhalte angepassten Prüfungsformaten in den letzten Jahren eine größere Bedeutung zugeteilt, da traditionelle Prüfungsformate die Aspekte ärztlicher Kompetenz nicht umfangreich testieren können (Fischer et al. 2010; Möltner et al. 2010). So heißt es auch in der Verordnung zur Neuregelung der zahnärztlichen Approbationsordnung:

> „Die Novellierung ist angesichts […] der veränderten Anforderungen an eine moderne und interdisziplinäre Lehre dringend erforderlich, um auch künftig die Qualität der zahnärztlichen Ausbildung als Voraussetzung für die zahnmedizinische Versorgung der Patientinnen und Patienten in einer älter werdenden Gesellschaft sicherzustellen.“ (Bundesrat 2017)

Bereits in einer Studie von Newble und Jaeger aus dem Jahr 1983 konnte nachgewiesen werden, dass eine Änderung von Prüfungsformaten und Bewertungsschwerpunkten gleichzeitig auch eine Veränderung im Lernverhalten der Studierenden hervorruft. Nach Etablierung eines neuen Bewertungsschemas für Studierende im letzten Semester stellten die Prüfer fest, dass sich die Studierenden in ihrem Lernverhalten auf das rein theoretische Abschlusstestat fokussierten, obwohl der Lehrschwerpunkt in der Studie auf klinische und praktische Aspekte des Kurses gesetzt worden war. Daraufhin wurden über einen Zeitraum von drei Jahren Daten aus einer Fragebogenbewertung bezüglich der Lerngewichtung von theoretischen und praktischen Aspekten gesammelt. In den Fragebögen wurde ermittelt, wie sich verschiedene Abschnitte der Prüfungsbewertung auf das Lernverhalten der Studierenden auswirkten. Diese Fragebögen beantworteten Studierende während des praktischen Jahres. Erst durch Umstellung auf eine stärkere Gewichtung der Benotung der klinischen Komponenten des Kurses fand auch ein Ausgleich der studentischen Lernaktivität zwischen klinischen und theoretischen Kursinhalten statt (Newble und Jaeger 1983).

Angelehnt an diese Studienergebnisse kommen Shumway et al. sowie Centeno et al. ebenfalls zur der Konklusion, dass die Lehrenden Prüfungen als ein bewusstes Instrument einsetzen sollten, um Studierende in ihrem Lernverhalten zu leiten und zu fördern (Shumway et al. 2003; Centeno et al. 2007). Auch Möltner et al. (2010) fordern eine verstärkte und bewusste Integration der Prüfungen als didaktisch eingesetztes Lehrmittel in der medizinischen Ausbildung.

Prüfungsformate können jedoch neben einem fördernden Effekt auch einen negativen Einfluss auf das Lernverhalten der Studierenden ausüben. Durch didaktisch falsch fokussierte Prüfungsformate oder Gewichtungen können die Lehrenden gegenteilige Effekte

in Bezug auf die Lernmotivation der Studierenden bewirken (Aebli 1997). Wenn Prüfungsformate ausschließlich auf einer reinen Reproduktion von Faktenwissen basieren, fördert das die Implementierung von sehr oberflächlichem Lernverhalten, anstatt dass Studierende aufgefordert sind, Transferleistungen zu erbringen (Newble und Jaeger 1983; Entwistle und Meyer 1992).

Epstein formulierte in seinem Review-Artikel folgende Leitpunkte, die zur Zielsetzung für jedes Prüfungsformat, in dem klinische Kompetenz bewertet wird, eingesetzt werden sollen. Zunächst dient eine Prüfungsevaluation dazu, die Fertigkeiten der Lernenden zu optimieren, indem sie Lernende motiviert und deren Lernverhalten dadurch positiv steuert. Prüfungen fördern durch eine gezielte Auswahl und Ausbildung der Studierenden die zukünftigen Ärzte, die durch suffiziente medizinische Kompetenzen wiederum das Gemeinschaftswohl sichern (Epstein 2007).

Prüfungen können helfen, das Curriculum in Bezug auf Schwächen der bisherigen Lehr- und Prüfungsmethoden zu untersuchen, indem sie erfassen, ob die Studierenden den Anforderungen an die spätere ärztliche Tätigkeit entsprechen. So können allgemeine Schwächen des curricularen Aufbaus frühzeitig erkannt und verbessert werden (Sopka et al. 2013).

Um den genauen Zusammenhang zwischen studentischem Lernverhalten und Prüfungsformaten zu eruieren, ist es ebenfalls sinnvoll, den Unterschied zwischen formativen und summativen Prüfungen zu erläutern.

1.1.1 Formative Prüfungen

Da die in diesem Forschungsvorhaben untersuchte Studie eine summative Prüfung darstellt, soll das formative Format nur der Vollständigkeit halber erwähnt werden.

In allen Definitionen wird das Konzept des Feedbacks als zentrales Grundelement der formativen Prüfungen genannt. Black und Wiliam definieren das formative Format als gesammelte Informationen sowohl von den Lehrenden als auch von den Studierenden innerhalb eines Lernprozesses, das zu einer Veränderung in Lehr- und Lernverhalten genutzt werden kann (Black und Wiliam 1998a).

Formative Prüfungen dienen dazu, den Studierenden eine konstruktive Rückmeldung über die bisher erbrachte Studienleistung zu erteilen und sie anzuregen, die fehlenden Lehrinhalte aufzuarbeiten (Sadler 1989). Laut Hudson liegt der Schwerpunkt dieses Formates auf einem selbstbestimmten Lernverhalten und wird nicht durch strikte curriculare Lehrpunkte fremdbestimmt, sodass Studierende ein tieferes Verständnis für wichtige Aspekte der klinischen Medizin entwickeln und diese in ihrer späteren Berufstätigkeit anwenden können (Hudson und Bristow 2006).

Durch das Fehlen einer Bestehensgrenze bei formativen Prüfungen kann die Lernmotivation der Studierenden gegenüber dem summativen Format jedoch deutlich reduziert sein (Raupach et al. 2013).

1.1.2 Summative Prüfungen

Summative Prüfungen testieren das erwartete Leistungsniveau und geben eine Bewertung der Kompetenzen, die die Studierenden vor Erreichen eines neuen curricularen Studienabschnittes aufweisen müssen (Van Der Vleuten 1996). Ein mit dem formativen Format vergleichbares explizites und konstruktives Feedback existiert bei diesem Prüfungsformat nicht. Die Studierenden erhalten ihr Feedback lediglich über Noten oder Prozentzahlen. Diese Punktzahlen werden genutzt, um zu ermitteln, ob Studierende die untersuchten Lehrziele erreicht haben (Sopka et al. 2013).

Durch ihre Funktion des direkten Vergleiches und die global verbreitete Anwendung dienen summative Prüfungen bisher zur Qualitätssicherung in der medizinischen und zahnmedizinischen Lehre (Shumway et al. 2003).

Raupach et al. untersuchten den Effekt zweier unterschiedlicher Lehrmethoden (konventionelle Vorlesungen vs. *Near-Peer*-Unterricht) auf die Prüfungsergebnisse von Medizinstudenten. Es konnte festgestellt werden, dass die Wahl der Lehrmethode nur einen geringen Effekt auf die Lernmotivation der Studierenden ausübte, jedoch die Wahl eines summativen Prüfungsformates in beiden Kohorten einen deutlicheren Lernanreiz bot als eine rein formative Prüfung (Raupach et al. 2010).

In einer weiteren Studie von Raupach et al. (2013) konnte nachgewiesen werden, dass summative Prüfungsformate einen objektiv deutlich größeren Lernanreiz bieten als explizit designte innovative Lehrformate. Ebenfalls wurde betont, dass sich Lehrende der Verantwortung bewusst sein müssen, die sie mit der Auswahl und Qualität der Prüfungsmethode auf das Lernverhalten der Studierenden ausüben.

1.2 Übersicht vorangegangener Studien bezüglich des Vergleiches von VSA- und SBA-Items

In Kapitel 1.1 wurde bereits auf die Kongruenz zwischen der Evaluationsform und dem studentischen Lehrverhalten hingewiesen, da für viele Studierende das explizite Studium der eingesetzten Prüfungstechnik im Vordergrund steht. In diesem strategischen Lernansatz versuchen Studierende ihre Prüfungsleistung nicht durch ein tieferes Verständnis des Lehrinhaltes zu verbessern, sondern durch gezieltes Einüben und Wiederholen der eingesetzten Prüfungsitems (McCoubrie 2004).

Basierend auf diesen Erkenntnissen etablierten Sam et al. 2016 eine Pilotstudie, in der 266 Medizinstudenten an einer formativen Prüfung teilnahmen. Das Very-Short-Answer-Format (VSA-Format) sollte als neues Frageninstrument eingeführt werden, um mit dem bereits bekannten Single-Best-Answer-Format (SBA-Format) verglichen zu werden. Beide Formate basierten auf gleichen Lehrinhalten, trotzdem erreichten alle 15 SBA-Items ein höheres Ergebnis als die ihnen gegenübergestellten VSA-Items ($P < 0{,}01$). Die Studie zeigt, dass die alleinige Verwendung von SBA-Items zu einem oberflächlichen Lernverhalten führen kann, das darauf abzielt, Assoziationen oder Hinweise durch Item-Distraktoren zu erkennen (Sam et al. 2016).

In einer weiteren Studie untersuchten Sam et al. (2018) das bereits eingeführte Prüfungsformat VSA in Bezug auf Reliabilität, Prüfungsleistungen und Akzeptanz. Darüber hinaus wurde ein neues Konzept zur maschinellen Auswertung der VSA-Items im Hinblick auf Kosten und Zeiteffizienz getestet. Erneut wurden 299 Medizinstudenten ausgewählt, um jeweils 60 Items im VSA- und SBA-Format zu bearbeiten. Es wurden zwei Kohorten gebildet, die jeweils beide Formate in unterschiedlicher Reihenfolge bearbeiteten. Die Auswertung erfolgte über eine App *(Practique, Fry),* die einem Server die Daten zusendete und anschließend die Levenshtein-Distanz anwandte, um die vorab genehmigten Lösungen mit den Antworten der Studierenden abzugleichen. Mit zusätzlicher Rater-Kontrolle dauerte die VSA-Auswertung im Mittel eine Minute und sechsunddreißig Sekunden. Abermals erzielten die Studierenden in allen SBA-Items höhere Punktzahlen. Die VSA-Items zeigten jedoch eine höhere Reliabilität (Cronbach's Alpha 0,91) sowie eine höhere punktbiseriale Korrelation ($P < 0{,}001$). Durch die Präsentation von Hinweisen unterlagen die SBA-Items einem signifikanten Cueing-Effekt (Sam et al. 2018).

Im folgenden Abschnitt werden die von Sam et al. untersuchten Prüfungsformate für ein tieferes Verständnis detaillierter vorgestellt. Zunächst wird das bereits etablierte SBA-Format eingehend beschrieben, anschließend wird auf das VSA-Format eingegangen.

1.2.1 Aufbau des Single-Best-Answer-Formates von Multiple-Choice-Items

Aufgrund ihrer hohen Objektivität und Reliabilität sowie ihrer Kosteneffizienz (Sam et al. 2016) werden sowohl im vorklinischen als auch im klinischen Abschnitt des Zahnmedizinstudiums neben mündlichen Prüfungen meist Multiple-Choice-Fragen (SBA) in Prüfungen zum Testen theoretischen Wissens angewandt.

Es handelt sich dabei um schriftliche Testate mit in sich geschlossen gestellten Fragen, die elektronisch über das universitätseigene „Digitale Prüfungs- und Schulungszentrum" (DiPS) ausgewertet werden.

Für jedes Item werden meist vier bis fünf verschiedene Antwortmöglichkeiten generiert, bei denen entweder genau eine (Single-Best-Answer) oder mehrere Antworten (Multiple-Choice) als korrekt gelten. Die Basis des Items beinhaltet die Frage, die beantwortet werden muss.

Zusätzlich zu der korrekten Antwortoption werden die alternativen Optionen (Distraktoren) gelistet, die den Prüflingen gleichzeitig präsentiert werden (Haladyna 2004).

Über die Anwendung von SBA-Items ist es mittels der Auswahl an Antwortoptionen möglich, in einem kurzen Zeitumfang einen großen Bereich aus theoretischem Wissen, Kompetenzen und Fähigkeiten abzufragen. In einem Item können Distraktoren aus unterschiedlichen Themenbereichen kombiniert werden (Gerhard-Szep et al. 2016). Die Ergebnisse einer Prüfung sollen einen Überblick über die erlernten Fähigkeiten und den Wissenstand des Prüflings geben (Rogers und Yang 1996).

Dieses Item-Format unterliegt jedoch der Kritik, dass die Items durch die Präsentation von Antwortmöglichkeiten auch ohne konkretes Wissen mittels Hinweisen oder durch Ausschluss der nicht sinnvoll erscheinenden Optionen beantwortet werden können.

Der Schweregrad einer Frage wird durch die Qualität der falschen Antwortmöglichkeiten bestimmt. Die Schwierigkeit bei Synthese von vier plausiblen Antworten, die nicht direkt als falsch erkannt werden können, ist vor allem beim Abfragen von Basiswissen als nicht trivial zu betrachten (Sam et al. 2018).

1.2.2 Der Einfluss von Test-Wiseness in SBA-Items

Die Anwendung von SBA-Items kann ein Lernverhalten fördern, das nur darauf abzielt, die richtige Antwort mittels Mustern oder Hinweisen wiedererkennen zu können und diese nicht selbstständig zu generieren beziehungsweise das benötigte Wissen aus dem Gedächtnis abzurufen (Epstein 2007). In einer Studie von Brozo et al. (1984) wurden 1220 Multiple-Choice-Items an zwei amerikanischen Colleges und drei Universitäten basierend auf den Test-Wiseness-Strategien von Millman et al. (Millman et al. 1965) untersucht. 44 % der Items wiesen einen Cueing-Effekt auf und sogar 70 % der Fragen konnten durch Anwendung einer der Test-Wiseness-Strategien beantwortet werden. Dieser Effekt konnte den Testteilnehmern dazu dienen, die richtige Antwort unter den Distraktoren zu erkennen oder Distraktoren aufgrund ihrer Konstruktion auszuschließen (Brozo et al. 1984).

Dieses Verhalten kann sogar dazu führen, dass der Prüfungserfolg auf Kosten eines tieferen Verständnisses der zu prüfenden Thematik nur kurzfristig optimiert wird und sich ein oberflächliches Lernverhalten der Studierenden verfestigt (Newble und Entwistle 1986; McCoubrie 2004).

Die Fähigkeit, subtile Hinweise in den Antwortoptionen der SBA-Items zu erkennen und diese zur korrekten Beantwortung des Items zu nutzen, wurde erstmals 1964 von Gibb als Test-Wiseness definiert. Weitere Untersuchungen konnten zusätzlich nachweisen, dass Cueing-Effekte die entscheidende Basis der Test-Wiseness bilden (Thoma und Köller 2018).

Bei suffizienter und korrekter Synthese von SBA-Items sollte es dennoch für Prüfungsteilnehmer nicht durch alleinige Anwendung des Cueing-Effekts möglich sein, die korrekte Antwort ohne theoretisches Testwissen zu erkennen (Haladyna 2004).

In einer Studie von 1989 errechnete Farley, dass ein geübter Prüfer für die Erstellung eines einzelnen SBA-Items im Durchschnitt eine Stunde benötigen würde (Farley 1989). Da das Lehrpersonal in der medizinischen Ausbildung hauptsächlich aus Ärzten besteht, die die Lehrfunktion neben der medizinischen Tätigkeit ausüben, ist der zeitliche Faktor zu Erstellung von Items sehr limitiert. Schulungen im Bereich der Prüfungsentwicklung finden derzeit noch keine große Anwendung (Downing und Haladyna 2006).

Der Cueing-Effekt kann außerdem in seiner Wirkungsweise unterschieden werden. Ein positiver Effekt tritt auf, wenn ein Hinweis im Item zur korrekten Beantwortung des Items führt; ein negativer Effekt führt dementsprechend zur falschen Lösung des Items (Schuwirth et al. 1996). In der hier untersuchten Studie wird ein positiver Cueing-Effekt gewertet, wenn trotz vorheriger inkorrekter Beantwortung des VSA-Items ein Hinweis zur richtigen Bewertung des SBA-Items führt (Schuwirth et al. 1996).

Zur mathematischen Erfassung des Cueing-Effekts eines Prüfungsformates erstellten Sam et al. (2019a) eine Formel, die ebenfalls in dieser Studie angewendet werden soll.

1.2.3 Aufbau des *very-short-answer*-Formates

Very-short-answer-Items bestehen nur aus einem Item-Stamm, der die zu beantwortende Frage formuliert. Auf diesen Item-Stamm muss die korrekte Antwort selbstständig generiert werden.

Im Gegensatz zum künstlichen Konstrukt mit vorgegebenen Antwortmöglichkeiten der SBA-Items müssen zukünftige Zahnärzte in einem realen Umfeld agieren, in dem sie von Patienten mit offenen Fragen konfrontiert werden, ohne dass ihnen zusätzlich noch eine Auswahl an unterschiedlichen Antwortmöglichkeiten präsentiert wird, aus der die korrekte Antwort ausgewählt werden kann. VSA-Items wirken dadurch deutlich realitätsnäher (Veloski et al. 1999).

VSA-Items weisen eine größere inhaltliche Validität auf, wenn die Fähigkeit untersucht wird, richtige Antworten selbstständig herzuleiten, anstatt sie nur durch Hinweise zu erkennen (Sam et al. 2016). Die Synthese einer Antwort beansprucht gegenüber der Wiedererkennung unterschiedliche kognitive Fähigkeiten, die das studentische testbezogene Lernverhalten verstärken und die Studenten ein tieferes Lernverständnis entwickeln lassen (Eagle und Leiter 1964; McConnell et al. 2015).

Für die Lehrenden bieten VSA-Items eine größere Flexibilität bei der Erstellung der Prüfungsitems, da der Fokus auf thematischer Relevanz liegt. Thematiken müssen nicht in einem Konstrukt entwickelt werden, auf das fünf plausible Antwortmöglichkeiten passen können. Dadurch kann das Testieren von Basiswissen und numerischem Wissen vereinfachter gestaltet werden (Damjanov et al. 1995; Fenderson et al. 1997).

Basierend auf der Varianz an falschen Antworten einer VSA-Frage kann bei der Auswertung ein detailliertes und spezifischeres Feedback bei inhaltlichen Differenzen erstellt werden, das

von den Lehrenden angesprochen oder zu curricularen Verbesserungen aufgegriffen werden kann. So können offensichtliche Lücken oder Unklarheiten des Curriculums anschließend bearbeitet werden. Zusätzlich bietet das VSA-Format Studierenden die Möglichkeit, ein tiefes Verständnis des Themas beziehungsweise den Umfang ihres erlernten Wissens nachzuweisen, indem sie alternative Antworten anbieten können, bei denen es sich gegebenenfalls um weitere plausible oder zutreffende Antworten handeln kann (Sam et al. 2018; 2019b).

Ein deutlicher Nachteil gegenüber den SBA-Fragen ist die Kosteneffektivität. Trotz der Implementierung computergestützter Auswertungssysteme (z. B. *Practique, Fry*) wird durch die zusätzliche Kontrolle der vorgenommenen Computer-Bewertungen durch einen Marker ein hoher zeitlicher Aufwand benötigt. In der Studie von Sam et al. benötigten die Marker zusätzliche 95 Minuten und 51 Sekunden für alle 60 VSA-Items bei 299 Studienteilnehmern, nachdem bereits 80,2 % der korrekten Antworten durch das System identifiziert wurden. Die SBA-Items hingegen konnten zu 100 % durch das System ausgewertet werden (Sam et al. 2018).

1.3 Grundlegende Testgütekriterien zur qualitativen Analyse medizinischer Prüfungsqualitäten

Zur Auswahl der richtigen Evaluationsmethode definierte van der Vleuten 1996 fünf Kriterien, um den adäquaten Einsatz einer Prüfungsmethode zu bewerten: die Reliabilität und Validität, den Einfluss auf das studentische Lernverhalten, die Akzeptanz durch Studierende und Lehrende sowie den Kostenfaktor.

Basierend auf den Studien von Sam et al. (2016, 2018, 2019a) fokussieren sich die Untersuchungsaspekte der hier vorgelegten Studie auf Reliabilität, Validität und deren Akzeptanz zur qualitativen Untersuchung beider Item-Formate. Zur besseren Verständlichkeit wird im Folgenden ein kurzer Überblick über die Aspekte der klassischen Testgütekriterien geboten.

1.3.1 Reliabilität

Die Reliabilität (R) kennzeichnet die Zuverlässigkeit und Aussagekraft eines Testes, mit der ein zu prüfendes Merkmal gemessen wird. Durch sie kann bestimmt werden, wie zuverlässig sich die Reproduzierbarkeit der gemessenen Ergebnisse eines Prüfungsformates unter gleichbleibenden Bedingungen darstellt (Chenot und Ehrhardt 2003).

Zur Bestimmung der Reliabilität kann auf vier Methoden zurückgegriffen werden. Je nach dem zu bestimmenden Merkmal eines Testes kann entweder die Retestreliabilität, die Paralleltestreliabilität, die Split-Half-Reliabilität oder die interne Konsistenz gemessen werden (Döring und Bortz 2016). In der vorliegenden Studie wurde die Reliabilität durch die interne Konsistenz mittels Cronbach's Alpha (α) berechnet.

Cronbach's Alpha bestimmt die Inter-Item-Korrelation mit einem Wert von null bis eins. Es werden die gemittelten Item-Zusammenhänge in Abhängigkeit zur Item-Anzahl eines Testes betrachtet. Da Cronbach's Alpha nur eine untere Grenze für die mathematisch bestimmte Reliabilität ($\alpha \leq R$) abbilden kann, wird in der Medizindidaktik für eine akzeptable Prüfung ein Mindestwert von $> 0{,}7$ angegeben. Bevorzugt werden Richtwerte für relevante Prüfungen ab 0,8 und 0,9 (Möltner et al. 2006; Pell et al. 2010).

Um zwischen „guten" und „schlechten" Prüfungsteilnehmern zu differenzieren (Chenot und Ehrhardt 2003), kann durch die Festlegung einer definierten Bestehensgrenze (60%-Regelung der UMG-Prüfungsverordnung) ein weiterer Aspekt der Reliabilität genutzt werden.

1.3.2 Validität

Die Validität zählt zu den klassischen Testgütekriterien und misst die Gültigkeit eines Testes (Wass et al. 2001). Durch sie wird geprüft, ob das Testmerkmal mit dem zu messenden Merkmal korreliert. Bei einer hohen Testvalidität können gezielte Prognosen auf Testergebnisse auch außerhalb der Testsituation getroffen werden. Grundsätzlich findet eine Einteilung in die vier Varianten der Inhalts-, Augenschein-, Konstrukt- und Kriteriumsvalidität statt (Moosbrugger und Kelava 2012).

Für ein genau zu bestimmendes Messkriterium ist die Validität klar definiert. In einer Prüfungssituation, in der es – abgesehen von der Bestehensgrenze – schwierig ist, die Definition eines „guten" oder „schlechten" Prüfungsergebnis zu formulieren, wird laut Möltner et al. (2006):

> „die Korrelation der bei einer Aufgabe erreichten Punktzahl mit diesem Kriterium als externe (Item-)Validität bezeichnet. […] Den Grad der Übereinstimmung von Aufgabe mit Gesamtpunktzahl bezeichnet man als Trennschärfe."

Mathematisch lässt sich die Trennschärfe (r) durch den Korrelationskoeffizienten nach Bravais-Pearson ermitteln. Auf das Szenario dieser Studie übertragen ist r dann definiert als Korrelation einer Antwort zur Punktesumme aller anderen Items. Ein Item hat eine hohe Trennschärfe, wenn Prüfungsteilnehmer mit einer hohen Gesamtpunktsumme viele Punkte in dem Item erreichen, hingegen „schlechte Prüfungsteilnehmer" dieses Item generell mit geringen Punktzahlen absolvieren. Dabei werden in der Literatur Trennschärfen-Werte von $> 0{,}3$ als gut bezeichnet und Werte von $\geq 0{,}2$ noch akzeptiert (Möltner et al. 2006).

1.4 Curriculare Unterschiede zwischen Zahn- und Humanmedizin

In den Studien vom Sam et al. (2018, 2019a) bezogen sich die Untersuchungen der Gütekriterien auf Kohorten bestehend aus Studierenden der Humanmedizin. Der größte curriculare Unterschied zwischen den Studiengängen Human- und Zahnmedizin besteht in dem deutlich größeren Praxisanteil des Studiums der Zahnmedizin. Durch die intensive und

kontrollierte Patientenbetreuung in vier der fünf klinischen Semester erhalten Absolventen bereits unmittelbar nach dem abschließenden elften Prüfungssemester (Regelstudienzeit), ohne anschließendes praktisches Jahr, die Approbation und können als Zahnarzt tätig werden (UMG Lehre 2020).

Schon in den vier klinischen Semestern der Untersuchungskurse sind Zahnmedizinstudenten aufgefordert, ihr theoretisches Wissen nicht ausschließlich durch Reproduktion in Testaten wiederzugeben, sondern müssen bereits den Transfer von der Theorie zur Praxis leisten.

Zum Ende des WS 2019/20 wurde dennoch im Kurs „Zahnersatzkunde II“ an der UMG eine summative Prüfung zur Leistungsevaluation angesetzt. Die theoretische Prüfung dieses fast ausschließlich klinisch-praktischen Kurses bestand standardmäßig aus 30 Multiple-Choice-Items (SBA-Items). Die Studierenden behandelten bereits im Kurs „Zahnersatzkunde I“ im vorangegangen Semester Kursinhalte und konnten außerdem in den letzten drei Semestern das theoretische Grundwissen im direkten Umgang mit Patienten anwenden. Diagnosen und Therapien mussten somit ohne die vorherige Präsentation einer korrekten Lösung sowie alternative Optionen getroffen werden.

Da dies einen der gravierendsten Kritikpunkte gegenüber SBA-Items darstellt (Veloski et al. 1999), ist es ein Anliegen der hier vorgelegten Studie zu untersuchen, ob Studierende in ihrem finalen Studienjahr durch bereits gesammelte praxisnahe Anwendungen vergleichbare Ergebnisse in den VSA-Items erzielen können.

1.5 Fragestellung und Zielsetzung

Angelehnt an die Studien von Sam et al. (2016, 2018, 2019a) soll das Prüfungswissen der Zahnmedizinstudierenden der UMG im finalen klinischen Jahr mittels Integration eines VSA-Formates genauer ermittelt werden. Ebenfalls werden SBA-Items, die bisher das standardisierte globale Prüfungsformat darstellen, den VSA-Items im direkten Vergleich mittels quantitativer statistischer Analyse gegenübergestellt.

Nach bestem Wissen liegt bisher keine Datenerhebung zum direkten Item-Vergleich zwischen VSA- und SBA-Format in Bezug auf die zahnmedizinische Lehre im finalen klinischen Jahr in Deutschland vor. In der Studie aus dem Bereich der Ausbildungsforschung sollen dementsprechend folgende Forschungsfragen beantwortet werden:

1. Inwiefern unterscheiden sich VSA-Items bezüglich Item-Schwierigkeit und Item-Trennschärfe von SBA-Items in summativen Prüfungen des zahnmedizinischen Studiums nach vorangegangenen klinischen Praxiserfahrungen und der Anwendung der vermittelten theoretischen Inhalte?
2. Welche Akzeptanz sowie welche Vor- und Nachteile ergeben sich aus den studentischen Selbsteinschätzungen im direkten Vergleich beider Formate?

Basierend auf den vorangegangenen Studien von Sam et al. (2016, 2018, 2019a) wurden folgende Hypothesen aufgestellt:

1. Die VSA-Items unterscheiden sich sowohl in der Item-Schwierigkeit als auch in der Item-Trennschärfe von SBA-Items. Die Diskrepanzen lassen sich in einem geringeren Ausmaß als in den Studien von Sam et. al nachweisen.
2. Durch die Vertrautheit der Studierenden mit den SBA-Items in den vorangegangenen vier Jahren des Studiums wird die Akzeptanz des VSA-Formates als sehr niedrig bewertet

2 Material und Methoden

Bei der hier vorgestellten Forschungsarbeit handelt es sich um den Vergleich zweier Prüfungsformate in einer prospektiven, nicht randomisierten Studie.

Es wurde nur eine Studienkohorte in die Studie eingeschlossen. Die 37 Studierenden der Zahnmedizin, die im Wintersemester 2019/20 am Kurs der Zahnersatzkunde II teilnahmen, fungierten als Probanden. Einschlusskriterium war die Einwilligung zur Studienteilnahme. Es lagen keine weiteren Ausschlusskriterien vor.

Die Datenerhebung erfolgte nach Abschluss des Kurses der Zahnersatzkunde II in zwei nacheinander ablaufenden summativen Prüfungen (à 30 Fragen) und einem abschließenden Fragebogen.

2.1 Genehmigung der Ethikkommission und Datenschutz

Das Ethikvotum der Ethikkommission der Universitätsmedizin Göttingen genehmigte den Ethikantrag am 05.09.2019 (Referenznummer 28/8/19). Es lagen keine ethischen oder rechtlichen Bedenken vor.

Jeder Studienteilnehmer wurde vor Prüfungsteilnahme mündlich und schriftlich über das Forschungsvorhaben und den Studienverlauf informiert. Vor der Erhebung der Daten wurde zusätzlich eine schriftliche Einwilligung zur Teilnahme eingeholt. Die Studienteilnahme erfolgte freiwillig und die Einverständniserklärung konnte zu jedem Zeitpunkt ohne Angabe von Gründen widerrufen werden.

Hingegen war die Teilnahme an beiden schriftlichen Prüfungsformaten zur Erfolgsevaluation des Kurses der Zahnersatzkunde II für alle Kursteilnehmer verpflichtend.

Beide Prüfungen wurden zunächst im Zusammenhang mit der Stud.IP-Kennung von den Studierenden bearbeitet. Die studienbezogene Auswertung zur statischen Datenanalyse erfolgte mit einem anonymisierten Datensatz, der keine persönlichen Informationen enthielt. Jedem Studienteilnehmer wurde randomisiert eine numerische Kennung für beide Prüfungsformate zugeordnet.

2.2 Studiendesign und Prüfungsaufbau

Die Studierenden bearbeiteten die zwei Prüfungsformate in elektronischer Form an separaten Computerbildschirmen im Digitalen Prüfungs- und Schulungszentrum (DiPS) der UMG.

Das Fragekonstrukt wurde so konzipiert, dass beide Prüfungsformate inhaltlich die identischen Thematiken behandelten. Für jeden Prüfungsabschnitt standen den Studierenden fünfundvierzig Minuten zur Verfügung. Es war für Studierende nicht möglich,

während eines Prüfungsabschnittes auf markierte Antworten oder Items des anderen Formates zurückzugreifen.

Damit durch längere Prüfungszeiten kein Nachteil entstand, wurde für jeden Studierenden das Prüfungsformat als Ergebnis gewertet, indem eine höhere Punktzahl erreicht wurde.

2.2.1 Itemaufbau im Very-Short-Answer-Prüfungsformat

In dem zu Beginn getesteten Prüfungsformat beantworteten die Studierenden die dreißig Fragen als offen gestellte VSA-Items, bei denen sie eigenständig formulierte, aus ein bis vier Worten bestehende Antworten generierten (Abbildung 1).

Die Items wurden über die Funktion „Freitext“ des Programms ItemManagementSystem (IMS) vom Umbrella Consortium for Assessment Networks (UCAN, Heidelberg, Deutschland) konzipiert.

Zur Bewertung der VSA-Items wurde pro Frage eine konkrete Anzahl an plausiblen Antwortoptionen festgelegt, die ein Prüfer in der abschließenden Prüfungsauswertung als korrekt markieren musste.

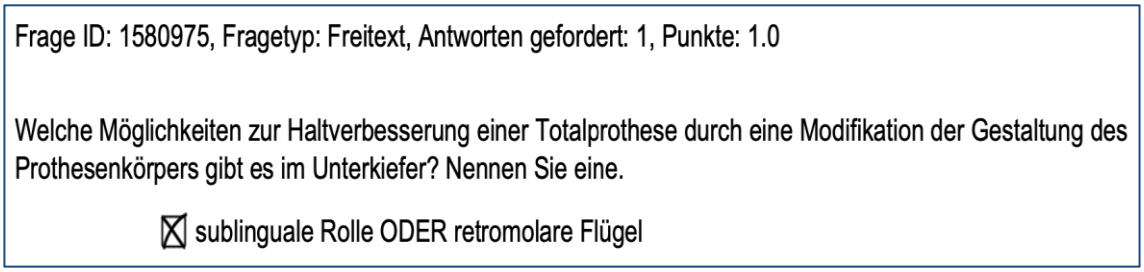


Abbildung 1: Beispielfrage im VSA-Format
Zur Beantwortung dieses Items muss eine der beiden Antwortoptionen durch den Studienteilnehmer benannt werden.

2.2.2 Item-Aufbau im Single-Best-Answer-Prüfungsformat

Im zweiten Prüfungsabschnitt bearbeiteten die Studierenden die 30 Fragen erneut, jedoch im SBA-Format. Die SBA-Items bestanden aus in sich geschlossen gestellten Fragen im Multiple-Choice-Format.

Jedes Item listete fünf Antwortoptionen, von denen nur eine Aussage als richtig gewertet wurde (siehe Abbildung 2).

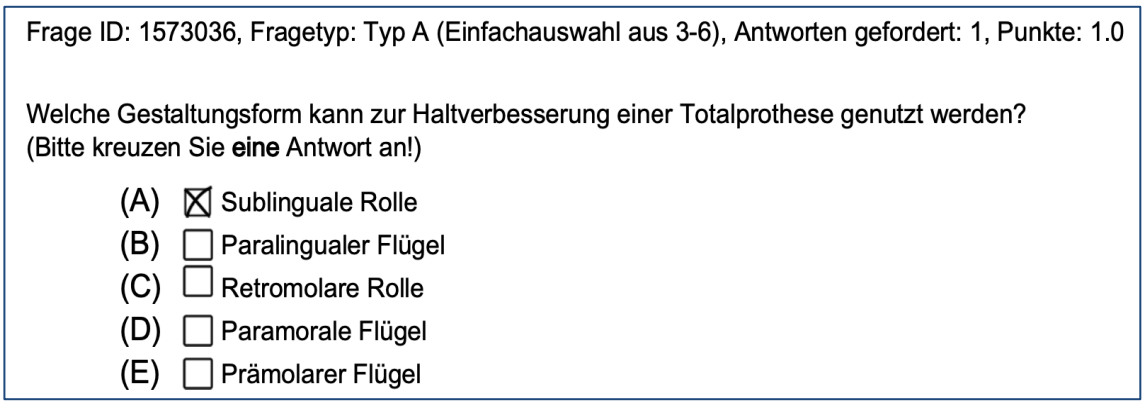


Abbildung 2: Beispielfrage im SBA-Format

Präsentation von fünf möglichen Antworten, von denen nur Antwortoption „A" als korrekt gewertet wurde.

Die SBA-Items stammten teils aus dem bereits im Programmsystem existierenden Fragenpool der UMG, teils pflegte der Prüfer sie über die Option „Typ A (Einfachauswahl aus 3-6)" in das IMS-Programm neu ein.

Die Prüfungen des Kurses der Zahnerhaltungskunde II aus vorherigen Semestern bestanden sämtlich aus reinen SBA-Formaten.

2.3 Elektronische Auswertungsstrategien der Prüfungsergebnisse

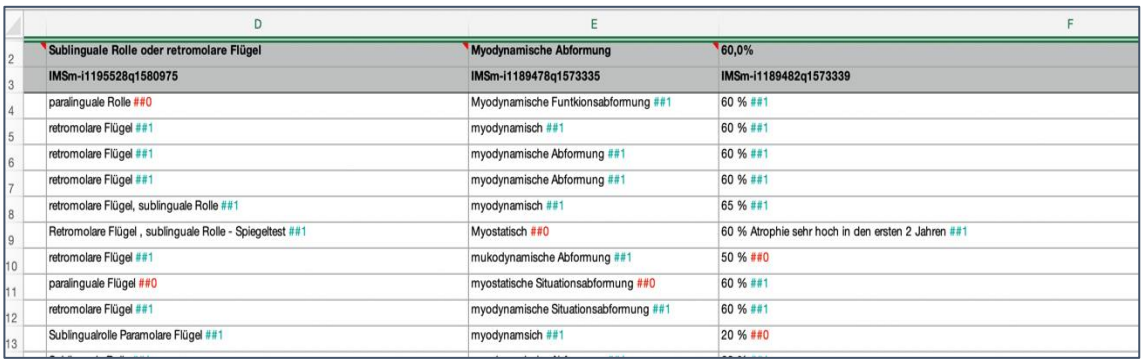
Sowohl die individuellen Ergebnisse der Studenten als auch die einzelnen Item-Parameter und Prüfungsdaten lagen dem Prüfer direkt nach Abschluss der Klausur zur statistischen Analyse vor.

Die SBA-Items wertete das IMS-Programm Examinator[2] direkt elektronisch aus.

Die deskriptiven Daten wurden mittels absoluter und relativer Häufigkeit sowie Mittelwert, Median und Standardabweichung dargestellt. Zudem berechnete das Programm für die Distraktorenanalyse die Verteilung und den Diskriminationsindex der einzelnen Items.

Für die Auswertung der VSA-Items musste ein Zwischenschritt eingefügt werden, indem ein Marker die von den Studierenden gegebenen Antworten in einer Excel-Tabelle einzeln auflistete und bewertete.

Spalte A der Tabelle listete alle Studienteilnehmer (n = 37) auf, denen jeweils eine Zeile mit den gelisteten Antworten zugeteilt wurde. Der Marker verglich die Aussagen mit den vorab festgelegten, gültigen Antwortmöglichkeiten und bewertete jede Aussage mit: „xx0" für falsche Antworten und „xx1" für richtige Antworten (siehe Abbildung 3).



	D	E	F
2	**Sublinguale Rolle oder retromolare Flügel**	**Myodynamische Abformung**	**60,0%**
3	**IMSm-i1195528q1580975**	**IMSm-i1189478q1573335**	**IMSm-i1189482q1573339**
4	paralinguale Rolle ##0	Myodynamische Funtkionsabformung ##1	60 % ##1
5	retromolare Flügel ##1	myodynamisch ##1	60 % ##1
6	retromolare Flügel ##1	myodynamische Abformung ##1	60 % ##1
7	retromolare Flügel ##1	myodynamische Abformung ##1	60 % ##1
8	retromolare Flügel, sublinguale Rolle ##1	myodynamisch ##1	65 % ##1
9	Retromolare Flügel , sublinguale Rolle - Spiegeltest ##1	Myostatisch ##0	60 % Atrophie sehr hoch in den ersten 2 Jahren ##1
10	retromolare Flügel ##1	mukodynamische Abformung ##1	50 % ##0
11	paralinguale Flügel ##0	myostatische Situationsabformung ##0	60 % ##1
12	retromolare Flügel ##1	myodynamische Situationsabformung ##1	60 % ##1
13	Sublingualrolle Paramolare Flügel ##1	myodynamsich ##1	20 % ##0

Abbildung 3: Auswertung der VSA-Fragen in Excel

Der Prüfer bewertet die Antworten aller Studienteilnehmer mit „##0“ für falsche und „##1“ für richtige Item-Antworten. Die akzeptierten Antworten sind in Zeile 2 gelistet, Zeile 3 listet die jeweilige Item-Bezeichnung. Jede Spalte führt sämtliche Antworten eines Items auf. Die Zeilen 4-40 stehen jeweils für die Antworten eines Studierenden.

Anschließend errechnete das Programm Examinator[2], basierend auf der Marker-Bewertung, die deskriptiven Daten der jeweiligen Items.

2.4 Inferenzstatistische Untersuchungen der Rohdaten

Die statistische Datenaufbereitung der Rohdaten erfolgte mittels des Statistikprogramms IBM SPSS Statistical Software Version 26.0 für Microsoft Windows (IBM, Armonk, New York, USA).

Zum Vergleich der Item-Schwierigkeiten zwischen der SBA- und VSA-Prüfung wurde eine Varianzanalyse mit Messwiederholung gerechnet. Die Voraussetzung der Homoskedastizität der Daten für Varianzanalysen wurde mit Mauchly’s Tests für Sphärizität überprüft. Bei signifikanten Abweichungen wurden die Freiheitsgrade der Varianzanalyse mit der Greenhouse-Geisser-Korrektur modifiziert.

Die Mittelwerte wurden zwischen der SBA- und VSA-Klausur mit abhängigen t-Tests verglichen. Bei allen Mittelwertvergleichen wurden die standardisierten Residuen abgespeichert und in Histogrammen auf Abweichungen von der Normalverteilung untersucht. Post-hoc t-Tests wurden mit der Bonferroni-Methode für multiple Vergleiche korrigiert.

Da visuelle Inspektionen der Residuen der Item-Schwierigkeiten der einzelnen Items auf teilweise starke Abweichungen von der Normalverteilung hindeuteten, wurden zum Vergleich der einzelnen Items zwischen den beiden Klausurtypen nicht-parametrische Wilcoxon-Vorzeichen-Rangtests berechnet.

Zur Untersuchung, ob die Bereitstellung von Antwortoptionen zu einer positiven Hinweisrate führt, wurden die Ergebnisse aus beiden Formaten gegeneinander geprüft. Eine positive Hinweisrate lag dann vor, wenn ein Studierender im VSA-Format eine inkorrekte Antwort gab, dieselbe Frage jedoch im SBA-Format korrekt bewertete (Sam et al. 2019a).

Die Berechnung der positiven Hinweisrate erfolgte mittels der von Sam et al. (2019a) entwickelten Formel:

$$\text{Positive Hinweisrate} = \frac{\textit{Anzahl der Studierenden,die VSA inkorrekt UND die SBA korrekt beantworteten}}{\textit{Anzahl der Studierenden,die die VSA inkorrekt beantworteten}} \times 100.$$

Die positive Hinweisrate wurde mit einem Einstichproben-t-Test gegen den Wert 20 % verglichen, da sich eine positive Hinweisrate von 20 % ergibt, wenn alle Studierenden, die ein VSA-Item falsch beantwortet haben, beim korrespondierenden SBA-Item lediglich geraten haben.

Als Maß für die interne Konsistenz wurde Cronbach's Alpha errechnet. Zur Prüfung, ob sich die interne Konsistenz erhöhen ließ, wenn Items mit extremen Item-Schwierigkeiten ausgeschlossen wurden, wurde die Reliabilitätsanalyse unter Ausschluss von fünf Items wiederholt.

Zu dem inferenzstatistischen Vergleich der Cronbach's Alpha-Werte des SBA- und VSA-Formates wurde ein F-Wert entsprechend des Bose- und Finney-Ansatzes berechnet (Feldt 1980).

2.5 Aufbau des Fragebogens zur Studieneinschätzung bezüglich Akzeptanz und Lernverhalten

Zur Evaluation der subjektiven Wahrnehmung beider Fragenformate beantworteten die Studierenden einen Evaluationsbogen im direkten Anschluss an die Prüfung. Es konnten 37 von 37 erfassten Datensätzen ausgewertet und zur Untersuchung verwendet werden.

Die Studierenden bewerteten die folgenden vier skalierten Aussagen auf einer Likertskala. Zur Auswertung wurden die Skalierungen „1:trifft voll zu" und „2: trifft eher zu" als Zustimmung gewertet; „4: trifft kaum zu" und „5: trifft nicht zu" galten als Verneinung der Aussage.

Die erste Frage lautete, ob Items im *single-best-answer*-Format („Multiple-Choice-Fragen") einfacher zu beantworten waren als Items im *very-short-answer*-Format („offen gestellte-Fragen").

Als zweites sollte die Aussage bewertet werden, ob das VSA-Format eine repräsentativere Darstellung davon bietet, wie Inhalte und Fragen in der klinischen Praxis beantworten werden müssen.

Die dritte Aussage lautete, dass Studierende für eine Klausur, die im reinen VSA-Format gestellt ist, ihre Lernstrategie und Vorbereitung ändern würden.

Zum Schluss sollten die Studierenden beantworten, ob sie den Gebrauch von VSA-Fragen gegenüber SBA-Fragen in Klausuren bevorzugen würden.

Im Anschluss an die Datenerhebung wurden die Ergebnisse des Fragebogens über die Software Excel Microsoft Office Professional Plus 2019 für Microsoft Windows (Microsoft Corporation, Redmont, Washington, USA) auf Datenblätter übertragen, grafisch aufbereitet und mittels prozentualer Verteilung miteinander verglichen.

Der Fragebogen ist in Anhang A angefügt.

3 Ergebnisse

Bei den Studierenden (n = 37) des vierten klinischen Semesters der UMG handelte es sich um 28 weibliche und neun männliche Studienteilnehmer.

Alle 37 Studierenden (100 %) nahmen an beiden Prüfungsformaten teil und absolvierten sämtliche Items in beiden Formaten (je 30 Items) in der vorgegebenen Zeit (100 %). Insgesamt konnten durch beide Formate 2220 Antworten erfasst werden (37 x 30 x 2). Die Item-Antworten aller Studienteilnehmer konnten in die Analyse eingeschlossen werden.

Bei den untersuchten Gütekriterien wird vorausgesetzt, dass die Item-Bewertung durch ein reines Punktesystem erfolgte, sodass eine Gesamtbewertung durch die Summenanzahl der Punkte möglich ist (Möltner et al. 2006).

3.1 Überblick über die Gesamtergebnisse des VSA- und SBA-Formates im direkten Vergleich

Bei visueller Inspektion der Gesamtpunkteverteilung (siehe Abbildung 4) zeigte sich eine deutlich größere Streuung der SBA-Werte gegenüber den VSA-Werten. Auffällige Datenausreißer lagen in beiden Datensätzen nicht vor.

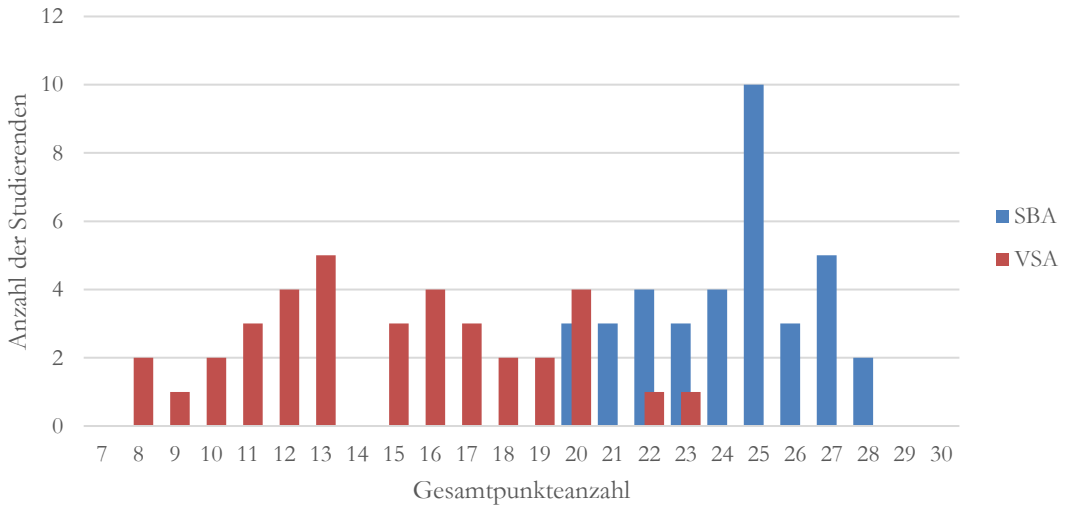


Abbildung 4: Häufigkeitsverteilung der erreichten Punktzahlen im Vergleich VSA- und SBA-Format
Für jedes korrekt markierte Item (n = 30) erhielten die 37 Studienteilnehmer einen Punkt, maximal konnten 30 Punkte pro Prüfungsformat erreicht werden. SBA: Single-Best-Answer-Items; VSA: Very-Short-Answer-Items.

Für jedes korrekt markierte Item erhielten die Studierenden einen Punkt, insgesamt konnten maximal dreißig Punkte je Prüfungsformat erzielt werden.

Die im Mittel erreichten Gesamtpunktzahlen beim VSA- und beim SBA-Format zeigten signifikante Unterschiede. Während im SBA-Format im Mittel 24,19 Punkte (max. 28,0

Punkte, min. 20,0 Punkte, SD = 2,31) erreicht wurden, erzielten die Studierenden im VSA-Format durchschnittlich nur 14,86 Punkte (max. 23,0 Punkte, min. 8,0 Punkte, SD = 3,97).

Die definierte Bestehensgrenze des SBA-Formates (Mittel-20 %) lag bei 18,0 Punkten und wurde von 37 (100 %) der Teilnehmer erreicht.

Im VSA-Format wurde die definierte Bestehensgrenze nach Anwendung der UMG-Gleitklausel (Mittel-20 %) auf 11,89 Punkte herabgesetzt. Damit bestanden 29 (78,38 %) der Prüfungsteilnehmer, acht (21,62 %) erreichten nicht die nötige Punktzahl zum Überschreiten der Bestehensgrenze.

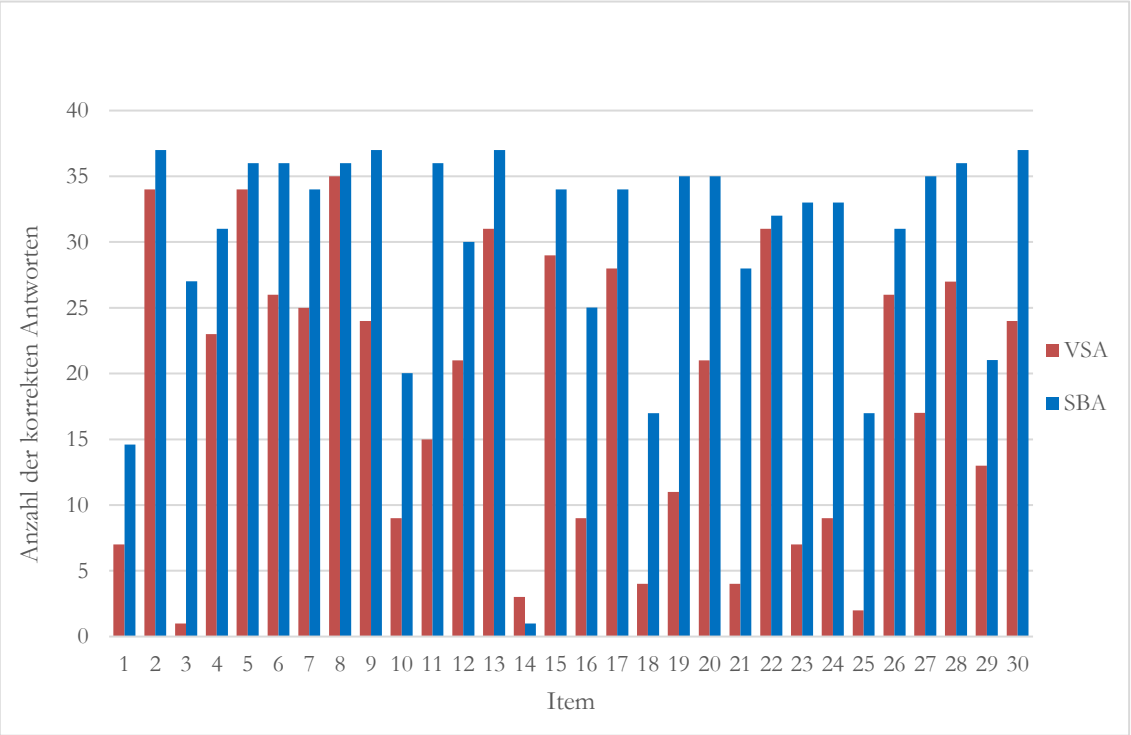


Abbildung 5: Anzahl korrekter Antworten pro Item im VSA- und SBA-Format
In zwei konsekutiven Prüfungen wurden mit je 30 Items (n = 60) Lehrinhalte des neunten Semesters geprüft. Im VSA-Format generierten die Teilnehmer ihre Antwort in ein bis fünf Worten selbst. Im SBA-Format wählten sie eine Antwort aus fünf vorgegebenen Optionen. Max. 37 korrekte Antworten je Item. SBA: Single-Best-Answer-Items; VSA: Very-Short-Answer-Items.

Werden die erreichten Gesamtpunktzahlen pro Item betrachtet (siehe Abbildung 5), so ist auffällig, dass das Items 14 häufiger im VSA-Format korrekt beantwortet wurden als im SBA-Format (siehe 3.2.1 Item-Schwierigkeit).

Auffällige Unterschiede bei der erreichten Punkteverteilung innerhalb der beiden Formate eines Items traten vor allem bei Item 3 auf, bei dem nur ein Proband die korrekte Antwort nennen konnte, jedoch wählten nach der Präsentation der Distraktoren 26 weitere Probanden die korrekte Lösung im SBA-Format aus. Ebenfalls auffällig sind Item 33, mit einer Differenz von 26 korrekten Antworten, die Items 19, 21 und 24, jeweils mit einer

Differenz von 24 korrekten Antworten, sowie Item 11 mit einer Differenz von 21 korrekten Antworten.

3.2 Psychometrieauswertung der Rohdaten

Im folgenden Kapitel werden die Ergebnisse der jeweiligen Testgütekriterien zur quantitativen Prüfungsanalyse detailliert aufegführt.

3.2.1 Auswertungsergebnisse der Item-Schwierigkeit

Nach Möltner et al. (2006) ist die Item-Schwierigkeit als mittlere erreichte Punktzahl bei einem Item definiert. Auf beide Prüfungsformate bezogen bedeutet dies, dass Item-Schwierigkeiten mit dem relativen Anteil von Probanden übereinstimmten, der ein Item im Sinne höherer Merkmalsausprägungen beantwortete.

Als Richtwert wird ein Wert zwischen 0,4 bis 0,8 empfohlen (Möltner et al. 2006). Demnach erfüllten nur sieben Items des VSA-Formates und zwölf Items des SBA-Formates dieses Kriterium.

Eine Varianzanalyse mit Messwiederholung mit den Innersubjektfaktoren „Item-Format" (SBA vs. VSA) und „Items" (Items 1 bis 30) und der Item-Schwierigkeit als abhängiger Variable zeigte signifikante Haupteffekte des Formates ($F_{(1;36)} = 113{,}08$, $P < 0{,}01$, $\eta = 0{,}76$), der Items ($F_{(14{,}22;511{,}98)} = 35{,}20$, $P < 0{,}01$, $\eta = 0{,}49$) und eine Interaktion von Format und Items ($F_{(13{,}38;481{,}49)} = 6{,}28$, $P < 0{,}01$, $\eta = 0{,}15$).

Wie bei der Analyse der Gesamtpunktzahl beschrieben, gab es beim VSA-Format (Item-Schwierigkeit: MW ± SD = 0,50 ± 0,13) signifikant weniger richtige Antworten als beim SBA-Format (Item-Schwierigkeit: MW ± SD = 0,81 ± 0,08). Interessanterweise variierte der Unterschied zwischen den Fragetypen je nach Item. Ein Vergleich der Item-Schwierigkeit einzelner Items ist in Tabelle 1 aufgeführt.

Tabelle 1: Signifikanzen der Item-Schwierigkeit im Vergleich von SBA- und VSA-Format

Itemnummer	SBA		VSA		Z-Wert	P-Wert
	MW	SD	MW	SD		
1	0,38	0,49	0,19	0,40	-1,70	0,09
2	1,00	0,00	0,92	0,28	-1,73	0,08
3	0,73	0,45	0,03	0,16	-5,10	**<0,01***
4	0,84	0,37	0,62	0,49	-2,00	0,05
5	0,97	0,16	0,92	0,28	-1,00	0,32
6	0,97	0,16	0,70	0,46	-2,89	**<0,01***
7	0,92	0,28	0,68	0,47	-2,50	**0,01**
8	0,97	0,16	0,95	0,23	-0,58	0,56
9	1,00	0,00	0,65	0,48	-3,61	**<0,01***
10	0,54	0,51	0,24	0,43	-2,52	**0,01**
11	0,97	0,16	0,41	0,50	-4,38	**<0,01***
12	0,81	0,40	0,57	0,50	-1,96	0,05
13	1,00	0,00	0,84	0,37	-2,45	**0,01**
14	0,03	0,16	0,08	0,28	-1,00	0,32
15	0,92	0,28	0,78	0,42	-1,51	0,13
16	0,68	0,47	0,24	0,43	-3,02	**<0,01**
17	0,92	0,28	0,76	0,43	-1,73	0,08
18	0,46	0,51	0,11	0,31	-3,15	**<0,01***
19	0,95	0,23	0,30	0,46	-4,54	**<0,01***
20	0,95	0,23	0,57	0,50	-3,50	**<0,01***
21	0,76	0,43	0,11	0,31	-4,71	**<0,01***
22	0,86	0,35	0,84	0,37	-0,33	0,74
23	0,89	0,31	0,19	0,40	-5,10	**<0,01***
24	0,89	0,31	0,24	0,43	-4,90	**<0,01***
25	0,46	0,51	0,05	0,23	-3,87	**<0,01***
26	0,84	0,37	0,70	0,46	-1,21	0,23
27	0,95	0,23	0,46	0,51	-4,03	**<0,01***
28	0,97	0,16	0,73	0,45	-2,71	**0,01**
29	0,57	0,50	0,35	0,48	-1,63	0,10
30	1,00	0,00	0,65	0,48	-3,61	**<0,01***
MW 1-30	0,81	0,08	0,50	0,13	-5,22	**<0,01**

*P-Wert = < 0,05 nach Bonferroni-Korrektur für 30 Tests. MW: Mittelwert; SD: Standardabweichung; VSA: Very-Short-Answer-Items; SBA: Single-Best-Answer-Items.

In Tabelle 1 werden trotz Verwendung eines nicht parametrischen Tests zum Vergleich der beiden Prüfungs-Formate Mittelwert anstelle des Medians verwendet, da der Median wegen der dichotomen Item-Formate wenig informativ ist. Um einer Alphafehlerkumulierung entgegenzuwirken wurde die Bonferroni-Korrektur über 30 Tests verwendet.

Gemittelt über die Probanden zeigte sich eine positive Korrelation zwischen den Item-Schwierigkeiten des SBA- und VSA-Formates ($r_{(28)} = 0{,}69$, $P < 0{,}01$; vgl. Abbildung 6). So ist anzunehmen, dass schwierigere Items des SBA-Formates tendenziell auch als schwierigere Items im VSA-Format galten (siehe Tabelle 1: Item 1, 4, 10, 18 und 25).

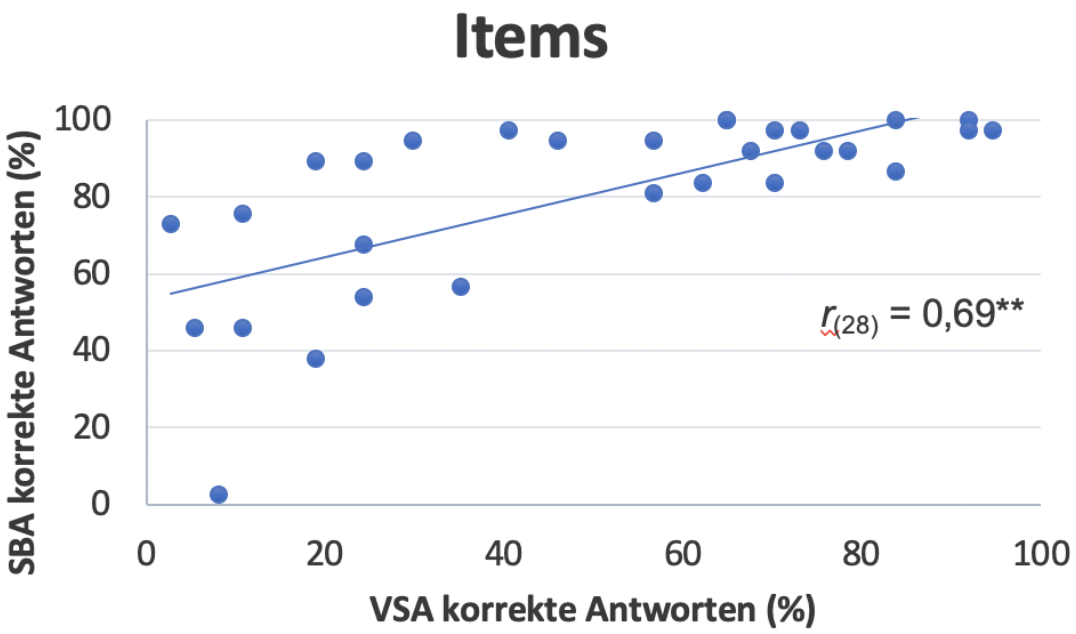


Abbildung 6: Korrelation der Item-Schwierigkeit im VSA- und SBA-Format gemittelt über die Probandenanzahl

VSA: Very-Short-Answer-Items; SBA: Single-Best-Answer-Items. Probandenzahl (n = 37)

In dieser Studie fand sich jedoch eine negative Korrelation der Item-Schwierigkeiten der SBA- und VSA-Formate, wenn über die Items gemittelt wurde ($r_{(35)}$ = -0.40, P = 0,01; vgl. Abbildung 7). Die negative Korrelation bedeutet, dass Studierende, die im SBA-Format gut abschnitten, tendenziell im VSA-Format ein schlechteres Ergebnis erzielten.

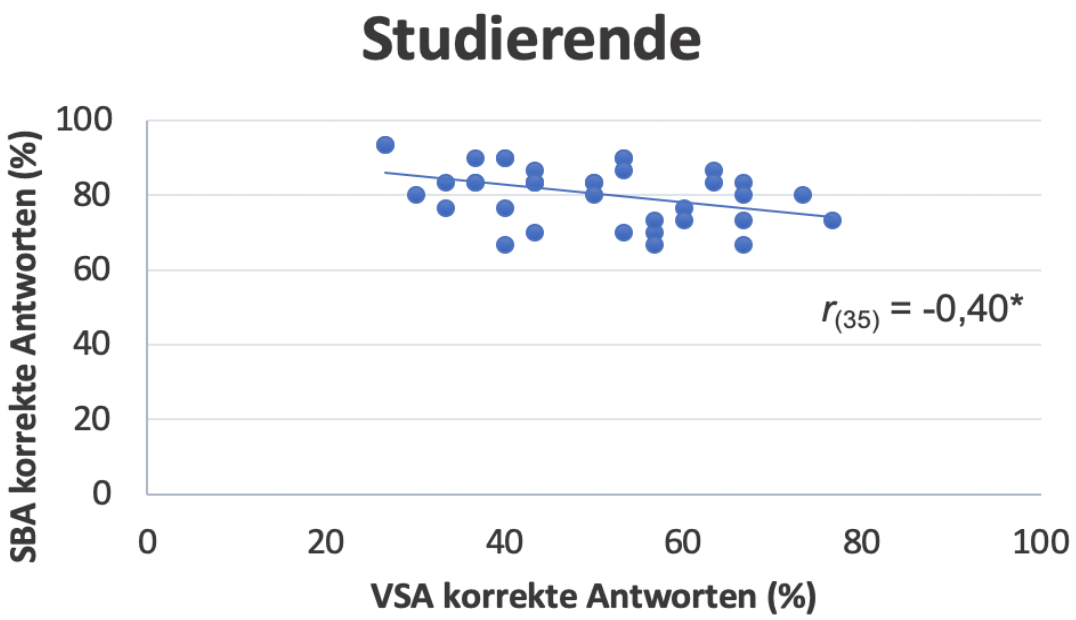


Abbildung 7: Korrelation zwischen der Item-Schwierigkeit im VSA- und SBA-Format gemittelt über die Items

Pro Item konnten max. 37 korrekte Antworten der Studienteilnehmer erfasst werden. Itemanzahl (n = 30) VSA: Very-Short-Answer-Items; SBA: Single-Best-Answer-Items.

3.2.2 Auswertungsergebnisse der Item-Trennschärfe

Die Item-Trennschärfe wird als „Grad der Übereinstimmung von Aufgabe mit Gesamtpunktzahl“ definiert (Möltner et al. 2006).

Die durchschnittliche Item-Trennschärfe war beim VSA-Format (MW = 0,22, SD = 0,17) signifikant größer als beim SBA-Format (MW = 0,12, SD = 0,13; $t_{(29)}$ = 2,50, P = 0,02).

Der Diskriminationsindex gilt als alternativer Indikator zur Trennschärfe. Er misst die Differenz zwischen der relativen Frequenz der Auswahlen und dem oberen sowie unteren Terzil der Gesamtpunkte der Probanden. Der durchschnittliche Diskriminationsindex aller Fragen des SBA-Formates betrug MW = 0,19 (SD = 0,16). Bei den Fragen des VSA-Formates zeigte sich ein signifikant besserer durchschnittlicher Diskriminationsindex aller Fragen von MW = 0,29 (SD = 0,20; $t_{(29)}$ = 2,23, P = 0,03).

3.2.3 Ergebnisse der positiven Hinweisrate

Die positive Hinweisrate im SBA-Format, berechnet nach Sam et al. (Sam et al. 2019a) (siehe Kapitel 2.4), betrug MW = 83,12 (SD = 24,09) und lag damit signifikant höher ($t_{(29)}$ = 14,35, P < 0,01) als die zu erwartende positive Hinweisrate von 20 %, wenn alle Studierenden, die ein VSA-Item falsch beantworteten, beim zugehörigen SBA-Item geraten hätten. Die positive Hinweisrate beim SBA-Format war stark positiv mit dem über die Items gemittelten

Anteil an korrekten Antworten im VSA-Format korreliert ($r_{(28)} = 0.71$, $P < 0{,}01$; vgl. Abbildung 8).

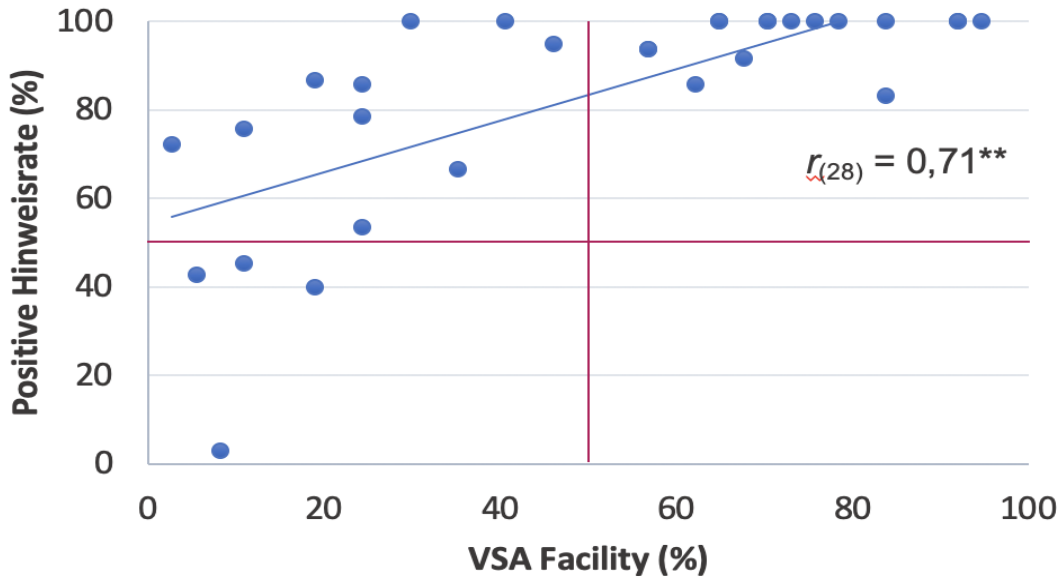


Abbildung 8: Korrelation der positiven Hinweisrate, gemittelt über die im VSA-Format korrekt genannten Items

Eine Positive Hinweisrate ergab sich aus dem Anteil von Studierenden, die Items im VSA-Format inkorrekt beantworteten, aber das SBA-Item korrekt markierten, dividiert durch die Anzahl der gesamten inkorrekten VSA-Antworten. Anschließend wurde mittels Einstichproben-t-Test gegen den Wert 20 % verglichen. Facility: Anzahl an korrekten Antworten; VSA: Very-Short-Answer-Items.

In Abbildung 8 lässt sich eine Unterteilung in vier Quadranten vornehmen. 33,3 % (n = 10/30) der Items liegen im linken oberen Quadranten. Diese Items wurden im VSA-Format als relativ schwer gekennzeichnet (VSA korrekte Antworten < 50 %) und wurden im SBA-Format deutlich besser beantwortet (Positive Hinweisrate > 50 %). Bei ihnen maskierten die SBA-Items eine unzureichende Wissensbasis, die mittels VSA-Items getestet wurde. Bedingt durch ein Wissensdefizit sowohl im SBA- und VSA-Format (VSA korrekte Antworten < 50 %) wiesen Items im linken unteren Quadranten (n = 4/30; 13,33 %) eine niedrigere positive Hinweisrate (< 50 %) auf. 40 % der Items lagen im oberen rechten Quadranten und wiesen eine hohe positive Hinweisrate (> 50 %) auf, kaschierten jedoch kein Wissensdefizit des VSA-Formates (VSA korrekte Antworten > 50 %).

3.2.4 Ermittlung der internen Konsistenz durch Cronbach's Alpha

Cronbach's Alpha (α) aller Items (n = 30) des SBA-Formates betrug 0,42. Um zu überprüfen, ob sich die interne Konsistenz der Prüfung dadurch erhöhen lässt, wenn Items mit extremer Item-Schwierigkeit ausgeschlossen werden, wurde die Reliabilitätsanalyse ohne fünf Items (vier Items mit Item-Schwierigkeit 1 und ein Item mit Item-Schwierigkeit 0,03) wiederholt. In dieser reduzierten Klausur (n = 25) erhöhte sich α jedoch nur geringfügig auf 0,42.

Das Cronbach's Alpha aller Items des VSA-Formates war mit 0,70 signifikant höher als das des SBA-Formates ($F_{(36;36)}$ = 2.05, P = 0,02). Beide Formate fielen jedoch unter den gewünschten Mindestwert an Reliabilität von > 0,80 (Möltner et al. 2006).

3.3 Evaluation der studentischen Akzeptanz und Auswirkungen auf das Lernverhalten

Der Fragebogen diente zur Beurteilung der Akzeptanz und Bewertung des neu eingeführten VSA-Formates durch subjektive Einschätzungen der Probanden. Die Studierenden bewerteten vier skalierte Aussagen auf einer Likertskala.

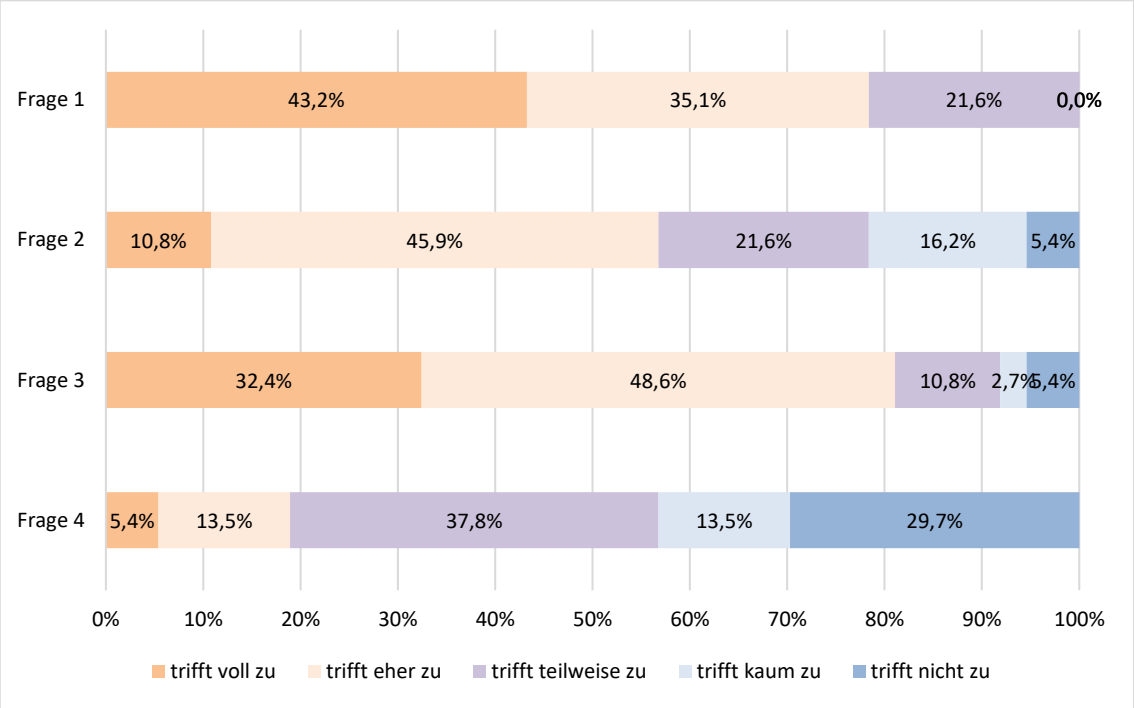


Abbildung 9: Auswertung der Probandenevaluation gegenüber der VSA-Akzeptanz
37 Studierende bewerteten vier skalierte Aussagen auf einer Likertskala. Zustimmungen wurden bei Bewertung mit „trifft voll zu" und „trifft eher zu" gewertet. Als Ablehnungen der Aussage wurde die Abstimmung mit „trifft kaum zu" und „trifft nicht zu" betrachtet.

1. Die Fragen im Single-Best-Answer-Format („Multiple-Choice-Fragen") waren für mich einfacher zu beantworten als die Fragen im Very-Short-Answer-Format („offen gestellte-Fragen").
2. Das Very-Short-Answer-Format bietet eine repräsentativere Darstellung davon, wie ich Inhalte und Fragen in der klinischen Praxis beantworten muss.
3. Für eine Klausur, die im reinen Very-Short-Answer-Format gestellt ist, würde ich meine Lernstrategie und Vorbereitung ändern.
4. Ich würde den Gebrauch von Very-Short-Answer-Fragen gegenüber Single-Best-Answer-Fragen in Klausuren bevorzugen.

Der Aussage, dass Items des SBA-Formates einfacher zu beantworten seien, stimmten 78,3 % der Probanden zu (vgl. Abbildung 9).

In Bezug auf die Authentizität empfanden 56,7 %, dass Items des VSA-Formates eine repräsentativere Darstellung bieten, wie theoretische Inhalte in der klinischen Praxis getestet werden. 21,6 % widersprachen dieser Aussage.

Die Veränderung der Lernstrategie erhielt große Zustimmung (81 %). Die Studierenden würden ihre Vorbereitung für Prüfungen anders fokussieren, wenn diese im reinen VSA-Format gestellt würden. Nur 8,1 % empfanden ihre vorab gewählte Lernstrategie als ausreichend.

Trotz repräsentativerer Darstellung der theoretischen Inhalte würden nur 18,9 % den Gebrauch von VSA-Items gegenüber SBA-Items bevorzugen. 43,2 % würden es bevorzugen, die Testierung durch SBA-Formate beizubehalten. Dabei ist jedoch anzumerken, dass die Studierenden sich im vorletzten Semester des Studiums befanden und bereits die vorangegangen acht Fachsemester durch SBA-Formate getestet wurden.

4 Diskussion

Die vorliegende Studie wurde durchgeführt, um die Integration eines neues Prüfungsformates (VSA-Items) in der zahnmedizinischen Lehre genauer zu betrachten. Dieses Format wurde mittels quantitativer Analyse den SBA-Items im direkten Vergleich gegenübergestellt. SBA-Items können Eindrücke über tatsächliche Kompetenz der Studierenden verfälschen, indem sie Wissensdefizite durch einen positiven Cueing-Effekt maskieren. Durch die Elimination von Distraktoren und Hinweisen minimieren VSA-Items dieses Risiko und bedingen andere kognitive Fähigkeiten zur Synthese der Item-Antwort. Unsere Ergebnisse unterstrichen die Vorteile bei Verwendung von VSA-Items als repräsentativere Darstellung des tatsächlichen Wissenstands in der zahnmedizinischen Lehre. Studierende erzielten im VSA-Format durchschnittlich 9,33 Punkten (31,1 %) weniger als im SBA-Format.

Ein weiterer Schwerpunkt der Studie lag auf dem Aspekt der Akzeptanzevaluation. In diesem Prozess wurde die Implementierung eines VSA-Formates seitens der Studierenden analysiert. Ein authentischeres Prüfungsformat wirkt sich sowohl auf ein tieferes Lernverständnis als auch auf die Lernmotivation aus. Tatsächlich empfanden nur 21,6 % der hier befragten Studienteilnehmer diese Aussage, VSA-Items sind ein realitätsnäheres Bewertungsinstrument (Frage 2, vgl. Abbildung 9), als nichtzutreffend. 81 % würden ihr Lernverhalten für ein VSA-Format in zukünftigen Prüfungen anpassen.

4.1 Diskussion der Methode

Das Ziel der Studie bestand nicht nur in einer rein qualitativen Analyse des VSA-Formates, sondern auch in einer genaueren Betrachtung des Formates in Bezug auf die Zahnmedizinische Lehre. In den klinischen Fächern der zahnmedizinischen Lehre wird bereits der Schwerpunkt auf praktische Tätigkeiten gelegt, sodass die Kompetenzbereiche aus den letzten vier Semestern direkt auf Patienten übertragen werden müssen. Um diesen Hauptaspekt der curricularen Unterschiede zwischen dem Zahn- und Humanstudium untersuchen zu können, wurde das vierte klinische Semester als Kohorte ausgewählt.

In dieser Studie wurde lediglich ein Prüfungsfach untersucht und nur Teilnehmer (n = 37) der UMG im vierten klinischen Semester Zahnmedizin eingeschlossen. Durch geringere Studierendenanzahlen im Vergleich zum Humanmedizin-Studium ist die Kohortengröße limitiert. Es lagen jedoch keine Datenausreißer vor, sodass alle erhobenen Datensätze zur statistischen Analyse eingeschlossen werden konnten. So konnte die Erfassung eines repräsentativeren Gesamteindruckes über die Studienkohorte gesichert werden.

Jegliche studentische Kennung wurde unmittelbar nach der elektronischen Auswertung zur Wahrung der Datenschutzrechte der Studierenden unkenntlich gemacht und sämtliche Daten in anonymisierter Form ausgewertet. Deshalb war es im Rahmen der

Akzeptanzevaluation nicht möglich, zu untersuchen, in wie weit die studentischen Bewertungen mit ihrem Ergebnis des VSA-Formats korrelierten.

Beide Item-Formate wurden zu Beginn des WS 2019/20 über die Kursleitung als verpflichtend festgelegt. Dieses bot eine repräsentative Darstellung über die Gesamtergebnisse der Abschlussevaluation. Negative Meinungen und Beurteilungen weniger motivierter Teilnehmer konnten ebenfalls erfasst werden.

In dieser Studie wurden 2220 Item-Antworten (30x37x2) in zwei elektronischen Prüfungen bewertet. Beide Item-Konstrukte behandelten inhaltlich identische Thematiken. Die Studierenden beantworteten die SBA-Items erst nach dem VSA-Format, um auszuschließen, dass VSA-Antworten durch zuvor präsentierte SBA-Distraktoren verfälscht wurden. Ähnlich des Studiendesigns von Sam et al. (2019a) stellte die Kombination beider Formate mit gleichem Inhalt auch eine Limitation dar. So kann die positive Hinweisrate dadurch verstärkt worden sein, dass Teilnehmer sich bereits mit dem Inhalt der VSA-Items auseinandergesetzt hatten und mittels der Antwortoptionen der SBA-Items die richtigen Schlüsse zogen.

An dieser Stelle sei darauf hingewiesen, dass es sich teilweise schwierig gestaltete, die zu prüfenden Thematiken in beiden Item-Formaten sinnvoll zu formulieren. Die Ergebnisse der Item-Schwierigkeit (MW ± SD = 0,81 ± 0,08) und Trennschärfe (MW = 0,12, SD = 0,13) des SBA-Formates zeigen, dass die Qualität der Items nur marginal der in der Literatur gewünschten Werte entspricht. Zwar konnte mittels Distraktoren eine vielfältigere Ausrichtung der Frage mit gleichzeitigen themenübergreifenden Feldern (Gerhard-Szep et al. 2016) erreicht werden, jedoch beeinflusste die Qualität der Distraktoren die Gütekriterien der Items (Sam et al. 2018). Konnte das zu prüfende Thema keine vier gleichwertigen qualitativen Distraktoren bieten, stieg die positive Hinweisrate und gleichzeitig erhöhte sich die Item-Schwierigkeit über einen Wert von 0,8. Diese Qualitätsminderung im SBA-Item-Format könnte von Studierenden durch bereits erlernte Strategien genutzt worden sein, um ihr Wissensdefizit im VSA-Format zu maskieren und ihre erreichte Punktzahl im SBA-Format zu steigern. Damjanov et al. (1995) und Fenderson et al. (1997) betonten hierzu den positiven Aspekt der VSA-Items, die ohne eine Konstruktplausibilität erstellt werden können und dem Lehrpersonal eine größere Flexibilität in der Prüfungsgestaltung bieten.

Ebenfalls lässt sich eine Verzerrung durch frühere Testungen in Bezug auf SBA-Items nicht komplett ausschließen. Bei Testierung von Basiswissen sind Prüfer bei der Item-Erstellung in ihren Distraktor-Optionen limitiert und Item-Inhalte können sich aus vorangegangenen Prüfungen überschneiden. Informationen über vorherige Testate werden von Studierenden an untere Semestergruppen weitergeleitet, so dass mögliche Vorkenntnisse über gewisse testierte Prüfungsinhalte bestehen können. Für diese Studie wurden jedoch ausschließlich neu generierte Items verwendet, sodass eine Verzerrung durch frühere Testungen als unwahrscheinlich gilt.

Sam et al. (2016, 2018, 2019a) führten die Prüfungen als formatives Format durch. In der Studie von Raupach et al. (2013) wurden keine signifikanten Unterschiede im Leistungserfolg der Studierenden zwischen einer Feedback-orientierten, formativen Prüfung (Black und Wiliam 1998) und einer summativen Prüfung verzeichnet. Die Lernmotivation seitens der Studierenden lässt sich jedoch durch ein summatives Prüfungsformat signifikant steigern. Basierend auf den Studienergebnissen ist das Prüfungskonzept in der hier vorgelegten Studie als summatives Format angesetzt worden.

Summative Prüfungen werden global eingesetzt und sollen die Qualität von Leistungsstandards in der medizinischen Lehre sichern. Sie kontrollieren erwartete Leistungsniveaus nach Erreichen eines festgelegten Lehrzieles und selektieren Studierende anhand ihres erreichten Kompetenzniveaus (Van Der Vleuten 1996; Shumway et al. 2003; Krasne et al. 2006; Sopka et al. 2013).

In summativen Formaten muss der ermittelte Leistungstand der Studierenden auch als authentisches Maß der zu prüfenden Kompetenz gelten können. Durch die Verwendung von SBA-Items mit hohen positiven Hinweisraten können Defizite im Wissensbereich verschleiert werden. Die durch die Studie gegenübergestellten niedrigeren VSA-Ergebnisse verwiesen auf Items, in denen ungenügende Kompetenzen verzeichnet werden konnten.

Die VSA-Auswertung zeigte, dass falsche Antworten oft kein Einzelfall waren, sondern gehäuft eine Vielzahl der gleichen falschen Antwort gegeben wurde. Sam et al. (2019b) verwies darauf, dass diese Problematik einen weiteren Vorteil bei der Verwendung von VSA-Items mit sich bringt. Durch die Prüfungsauswertungen können Lehrthemen und inhaltliche Aspekte erkannt werden, die zur Verbesserung der curricularen Lehre genutzt werden sollten.

Die Generalisierbarkeit ist möglicherweise durch die gleichen Lehrbedingungen der Kohorte eingeschränkt, um auf das gesamte zahnmedizinische Studium in Deutschland bezogen zu werden. Um eine größere Strichprobe und ein breiteres Spektrum an verschiedenen Lehrbedingungen zu erhalten, sollten weitere Kohorten verschiedener Universitäten auf die gleichen Aspekte untersucht werden. Ebenfalls sollten weitere Kohorten zu unterschiedlichen Zeitpunkten innerhalb des Studiums gewählt werden.

4.2 Diskussion der Ergebnisse

Bisher wurde in der medizinischen Lehre der Schwerpunkt von Prüfungen auf die Verwendung von SBA-Items gesetzt. Bereits Sam et al. (2016) verwiesen darauf, dass qualitative SBA-Items als Bewertungsinstrument einige Vorteile aufweisen. Neben einer hohen Reliabilität und Akzeptanz stellen sie ein zuverlässiges und kosteneffektives Bewertungsinstrument dar (Wass et al. 2001).

Das Lernverhalten kann durch gewählte Bewertungsmethoden gezielt gefördert werden, wenn sie vom Lehrpersonal richtig in Abstimmung auf das zu prüfende Merkmal eingesetzt

werden. Jedoch optimieren sich oft nur die Kompetenzen, die zur Leistungssteigerung im jeweiligen Prüfungsformat fördern. Dieses Verhalten konnte bereits in zahlreichen Studien (Sam et al. 2016, 2018, 2019a; Wass et al. 2001; McCoubrie 2004) beobachtet werden. Laut Wood (2009) sollte ein ideales Bewertungsinstrument mit den curricularen Anforderungen so abgestimmt werden, dass es immer ein tieferes Lernverhalten fördert.

Die untersuchte Studienkohorte erzielte in 29 der 30 SBA-Items höhere Punktzahlen (MW + 9,33 Punkte) als im VSA-Format, obwohl beide Formate dieselbe Wissensbasis testierten. Die Studienergebnisse bestätigen die Aussagen der oben aufgeführten Studien. So führte ein strategisches Erkennen von Hinweisen und Assoziationen zu einem oberflächlichen Lernverhalten. Mittels UMG Gleitklausel korrigierte sich die Bestehensgrenze von 16 benötigten Punkten auf 11,89 Punkte. Bei Betrachtung der reinen VSA-Resultate erreichten 21,62% (n = 8) dennoch nicht die benötigte Punktzahl zum Bestehen. Ohne Anwendung der Gleitklausel wäre die Bestehensgrenze lediglich von 21,62 % (n = 8) der Studienteilnehmer erreicht worden. Auffällig ist, dass im SBA-Format 100 % der Studienteilnehmer die benötigte Punktzahl mit einem Minimum an 20 Punkten überschritten.

Die Unterschiede in den Formaten sind dahingehend bedenklich, dass in der medizinischen Lehre fast ausschließlich SBA-Items in summativen Prüfungen einsetzt werden. Die hier erhobenen Daten implizieren, dass bei Prüfungen des Kurses Zahnersatzkunde II der UMG nicht die Kompetenz gemessen wird, die gemessen werden soll.

Aus Abbildung 8 (siehe Kapitel 3.2.3) lässt sich ableiten: je höher der Schweregrad der Items getestet wurde, desto geringer fiel die positive Hinweisrate aus. Items, die bereits von vielen Teilnehmern im VSA-Format beantwortet werden konnten, führten im SBA-Format zur zusätzlich korrekten Bewertung derjenigen, die zuvor die VSA-Items inkorrekt bearbeiteten. Daraus lässt sich folgern, dass Studierende bei leichteren Items deutlicher von der positiven Hinweisrate profitierten.

Bei Untersuchung der positiven Hinweisrate konnte jedoch ein deutlicher Unterschied zur Studie von Sam et al. (2019a) diagnostiziert werden. Im direkten Vergleich beider Streudiagramme (siehe Abbildung 10) ist die unterschiedliche Verteilung der positiven Hinweisrate deutlich erkennbar.

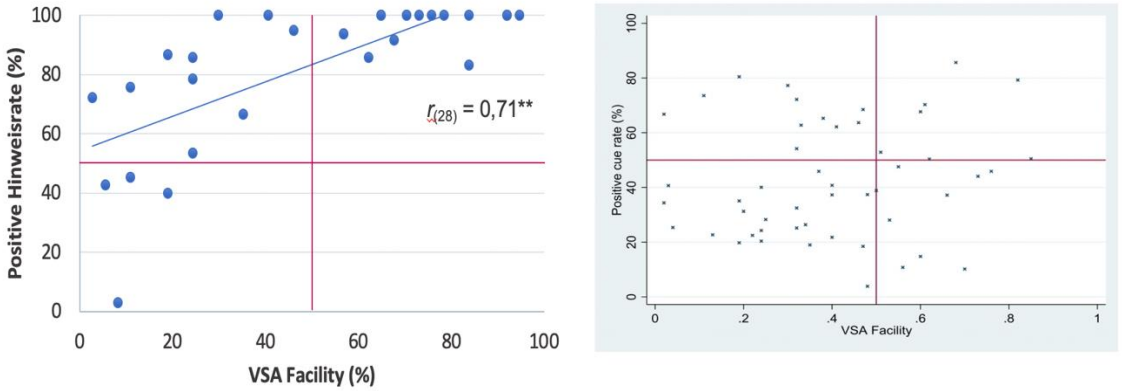


Abbildung 10: Vergleich beider Streudiagramme zur Korrelation positiver Hinweisraten gemittelt über Items mit korrekten Antworten

Zur Übersichtlichkeit wurden beide Diagramme gegenübergestellt und in Quadranten unterteilt. Das Streudiagramm von Sam et al. wird auf der rechten Seite aufgeführt, das Diagramm der linken Seite entspricht Abbildung 8 (Kapitel 3.2.3). Die von Sam et al. genannte „VSA-Facility" entspricht dem Anteil an Studenten, die Items korrekt beantworteten. VSA: Very-Short-Answer. (Sam et al. 2019a). Die Verwendung erfolgt mit freundlicher Genehmigung von BMJ Open.

Während sich 64 % der Items in dem Diagramm von Sam et al. (2019a) in den unteren beiden Quadranten (positive Hinweisrate < 50 %) lokalisierten, lagen in der hier untersuchten Forschungsarbeit 73 % der Items oberhalb der positiven Hinweisrate von > 50 %.

Die Verteilung der Daten aus der Studie von Sam et al. (2019a) ist deutlich homogener als die Verteilung der erhobenen Stichprobe. Die Distraktorenqualität der Items (n = 8) aus dem rechten unteren Quadranten (*VSA Facility* > 0,5; *positiv cue rate* < 50 %) führte bei Sam et al. dazu, dass Studierende Items korrekt im VSA-Format beantworteten, jedoch im SBA-Format die inkorrekte Lösung markierten. Die 48 % der Items aus dem unteren linken Quadranten demonstrierten einen unzureichenden Wissensstand innerhalb der Studienkohorte. Ein Großteil der Studienteilnehmer konnte diese Items weder im VSA- noch SBA-Format beantworten.

Dem gegenübergestellt waren die Werte aus der hier untersuchten Studie deutlich schiefer verteilt. Kein Item erreichte Werte im unteren rechten Quadranten und nur 13,33 % der Items lagen im unteren linken Quadranten. Dies mag darauf zurückzuführen sein, dass die Studienkohorte im Gesamten deutlich bessere Werte in Bezug auf die Itemschwierigkeit erzielte. So erreichte das SBA-Formate Werte von 0,81 und das VSA-Format 0,50, bei Sam et al. (2019b) lagen diese Werte bei 0,61 und 0,40.

Es lässt sich derzeit nur vermuten, welche Faktoren die fast doppelt so große positive Hinweisrate (83,12 %) im Vergleich zu den 42,7 % von Sam et al. (2019a) bedingt.

Der Mangel an klinischen Erfahrungen der testierten Studienkohorte, wurde von Sam et al. als Limitation ihrer Studie beschrieben. Unsere Kohortenauswahl umfasste diesen

Erfahrungsbereich. Es sollte untersucht werden, ob eine praktische Anwendung der Lehrinhalte zu einem differenten Ergebnis der VSA-Items führen würde, da sie laut Sam et al. von 69,2 % als authentischer und dem klinischen Alltag realitätsnäher empfunden wurden (2019a). Die Ergebnisse der untersuchten Stichprobe wiesen höhere positive Hinweisraten auf. Jedoch ist nicht ersichtlich, ob die Prüfungsformate in der zahnmedizinischen Lehre im Gesamten zu leicht formuliert waren, oder ob die Studienkohorte durch ein größeres Wissensspektrum höhere positive Hinweisraten und Klausurergebnisse erzielen konnten.

Die Studierenden der Zahnmedizin erreichten im Vergleich zur Vorstudie von Sam et al. (2016, 2019a) bessere Werte im VSA-Format. Jedoch wiesen die Prüfungen höhere Leistungsdifferenzen als die untersuchten Kohorten von Sam et al (2019a) auf. Diesbezüglich kann die aufgestellte Hypothese, dass Studierende nach praktischer Anwendung der Lehrinhalte bessere Ergebnisse in den VSA-Items erzielen würden, nur teilweise bestätigt werden. Es benötigt weiterführende Untersuchungen, um zu eruieren, ob die höheren Ergebnisse durch die Distraktorenqualität oder tatsächlich durch eine verstärke praktische Anwendung der Lehrinhalte bedingt sind.

Basierend auf der von Newble und Jaeger (1983) erstellten Forschungsarbeit, lässt sich durch die besseren SBA-Ergebnisse bei der Studienkohorte ebenfalls erkennen, dass das Lernverhalten trotz praktischer Orientierung eines Kurses durch ein theoretisches Prüfungsformat dominiert wird. Zum Ausgleich des Lernverständnisses bedarf es eine Umgewichtung der Bewertungsschwerpunkte.

Ein weiterer Aspekt der vorgelegten Studie galt der Untersuchung, ob die Implementierung eines neuen Bewertungsinstrumentes die Rahmenbedingungen der Wissensevaluation verbessern würde. Dabei wurden die Schwerpunkte auf Untersuchung der Reliabilität, Validität und Diskrimination der VSA-Items gesetzt.

Den Forschungsergebnissen von Sam et al. (2019a) entsprechend, demonstrierten die VSA-Items mit einem Cronbach's Alpha von 0,7 eine signifikant höhere Reliabilität als SBA-Items ($F_{(36;36)} = 2.05$, $P = 0{,}02$). Beide Formate blieben jedoch unter der in der Literatur geforderten Mindestwerte (> 0,8) für Reliabilitätskoeffizienten (Möltner et al. 2006). VSA-Items erwiesen sich folglich als reliableres Bewertungsinstrument zum Messen tieferer Lernverständnisse. Ebenfalls deckten sich die Resultate in Bezug auf akzeptablere Diskriminations- und Validitätswerte der VSA-Items. Das VSA-Format (MW = 0,22, SD = 0,17) erreichte signifikant größere Item-Trennschärfen als beim SBA-Format (MW = 0,12, SD = 0,13; $t_{(29)} = 2{,}50$, $P = 0{,}02$).

Die Studierenden der Zahnmedizin erreichten im Vergleich zur Studie von Sam et al. höhere Werte im VSA-Format. Jedoch wiesen die Prüfungen höhere Leistungsdifferenzen als die untersuchten Kohorten von Sam et al (2019a) auf. Diesbezüglich kann die aufgestellte Hypothese, dass Studierende nach praktischer Anwendung der Lehrinhalte bessere Ergebnisse in den VSA-Items erzielen würden, nur teilweise bestätigt werden.

Es benötigt weitere Stichprobenkontrollen in der zahnmedizinischen Lehre um zu eruieren, ob die höheren Ergebnisse der positiven Hinweisraten durch die Distraktorenqualität oder tatsächlich durch eine verstärke praxisnahe Anwendung der Lehrinhalte bedingt sind. Wünschenswert wäre es außerdem zu testen, wie sich Prüfungsdifferenzen und die Akzeptanz der VSA-Items bei Studienkohorten im beginnenden Studiensemester verhalten würden.

4.3 Schlussfolgerungen

Es lassen sich verschiebende Stärken und Limitationen in der dargestellten Studie identifizieren, jedoch deckten sich die Forschungsergebnisse mit den Schlussfolgerungen aus den Pilotstudien von Sam et al. (2016, 2018, 2019a) im weitesten Umfang. So fungieren VSA-Items in der zahnmedizinischen Lehre als akkurateres Bewertungsinstrument um Kompetenzen zu messen, mit denen Wissen auf Basis tieferer kognitiver Prozesse wiedergegeben wird, anstatt durch Erkennen von Hinweisen und Alternativen.

Sie demonstrierten im Mittel höhere Reliabilität und Validität als SBA-Items. Hingegen wiesen SBA-Items signifikant hohe Korrelationen mit positiven Hinweisraten auf. 78,3 % bewerteten die SBA-Items als deutlich leichter, jedoch empfanden sie nur 21,6 % der Studienteilnehmer als authentischer. Obwohl VSA-Items eine repräsentative Darstellung boten, wie Lehrinhalte im Umgang mit Patienten angewandt werden, würde laut 81 % der Befragten die Verwendung von VSA-Items, als alleiniges Prüfungs-Format, zur Umstellung im Lernverhalten führen.

Eine zukünftige Integration der VSA-Items als Bewertungsinstrument sollte daher als sinnvoll und notwendig betrachtet werden, um eine Optimierung in der zahnmedizinischen Lehre an der UMG zu erzielen. Jedoch sollte die Implementierung bereits in früheren Stadien des Studiums begonnen werden, um zu verhindern, dass sich ein oberflächlicheres Lernverhalten manifestiert.

Dennoch muss betont werden, dass es kein alleinstehendes ideales Prüfungsformat gibt, mit dem alle Facetten des medizinischen Kompetenzbereichs abgedeckt werden können. Van der Vleuten et al. (1996) verwiesen bereits darauf, den Fokus vielmehr darauf zu setzten verschiedenen Prüfungsformate zu kombinieren, um ein weites Feld zur Evaluation und Testierung der Studierenden zu erfassen.

5 Zusammenfassung

Single-Best-Answer-Items (SBA) gelten derzeit als das am weitesten verbreitete Bewertungsinstrument in der medizinischen Lehre. Das Format unterliegt jedoch der Kritik, ein oberflächliches Lernverhalten zu fördern und Hinweise auf Item-Antworten zu präsentieren. Angelehnt an die Studien von Sam et al. sollte durch die vorliegende Studie die Implementierung eines alternativen Prüfungsformates durch Very-Short-Answer-Items (VSA) getestet werden. Die Forschungsergebnisse von Sam et al. sollten mit Resultaten aus der zahnmedizinischen Lehre verglichen werden, um die Notwendigkeit einer Prüfungsoptimierung in der Zahnmedizin zu kontrollieren. Des Weiteren bewerteten Studierende VSA-Items in Bezug auf Authentizität und Akzeptanz.

In zwei inhaltlich identischen, summativen Prüfungen bearbeitete die Studienkohorte (n = 37) konsekutiv 30 Items, zunächst im VSA-Format und anschließend im SBA-Format. Den Studierenden wurde im Anschluss ein Fragebogen zur Akzeptanz, Authentizität und ihrem Lernverhalten vorgelegt. Nach statistischer Auswertung der deskriptiven Daten wurden die Ergebnisse beider Prüfungsformate gegenübergestellt. Zum Vergleich der Item-Schwierigkeit wurde eine Varianzanalyse und post-hoc Wilcoxon-Vorzeichen-Rangtests durchgeführt. Mittels Einstichproben-t-Test gegen einen Wert von 20 % wurden überprüft, ob die positive Hinweisrate signifikant vom Zufallsniveau abwich. Cronbach's Alpha wurde als Maß der Reliabilität bestimmt und zwischen den Formaten verglichen. Zusätzlich wurden die Trennschärfen der Items verglichen und Korrelationsanalysen durchgeführt.

Die Prüfung im VSA-Format bot in der Zahnmedizinischen Lehre eine akkuratere Präsentation der studentischen Wissensbasis. Die Ergebnisse demonstrierten signifikante Unterschiede in Bezug auf die Item-Schwierigkeit ($t_{(36)} = 10{,}63$, $P < 0{,}01$) zwischen beiden Formaten und entsprachen den Ergebnissen der Studie von Sam et al. im weitesten Umfang. Die untersuchte Studienkohorte erzielte in 29 der 30 SBA-Items höhere Punktzahlen (MW + 9,33 Punkte) als im VSA-Format, obwohl beide Formate dieselbe Wissensbasis testierten. VSA-Items zeigten mit einem Cronbach's Alpha von 0,7 eine signifikant höhere Reliabilität als SBA-Items ($F_{(36;36)} = 2.05$, $P = 0{,}02$). Im Mittel wiesen die VSA-Items eine signifikant bessere Item-Trennschärfe auf (MW = 0,22, SD = 0,17) als SBA-Items (MW = 0,12, SD = 0,13; $t_{(29)} = 2{,}50$, $P = 0{,}02$). Ebenfalls wurden sie von 56,7 % als authentischer empfunden. Die Akzeptanz gegenüber VSA-Formaten lag nach Konditionierung des Lernverhaltens durch SBA-Items nur bei 18,9 %. Die SBA-Items wiesen im Vergleich zu Sam et al. einen verdoppelten Wert (83,12 %) durch positive Hinweisraten auf. In der vorliegenden Studie konnten frühere Ergebnisse zur Auswirkung eines Bewertungsinstrumentes auf das Lernverhalten der Studierenden repliziert werden. Eine zukünftige Integration der VSA-Items in summative Prüfungen sollte als sinnvoll und notwendig betrachtet werden, um eine Optimierung in der zahnmedizinischen Lehre an der Universitätsmedizin Göttingen zu erzielen.

6 Anhang

Anhang A: Fragebogen zur Evaluation der studentischen Akzeptanz

Universitätsklinikum Medizinische Fakultät

Studienleiter und Ansprechpartner:
Prof. Dr. med. T. Raupach, MME
Abteilung Kardiologie & Pneumologie
Universitätsmedizin Göttingen
Robert-Koch-Straße 40, 37075 Göttingen
Telefon: 0551/398922
Mail: raupach@med.uni-goettingen.de

Probanden/innen-Befragung zur Studie:

„Vergleich der SBA- und VSA-Frageformate in der zahnmedizinischen Lehre hinsichtlich des Schweregrades, der Trennschärfe und deren Akzeptanz."

Sehr geehrte Studentin, sehr geehrter Student,

im Folgenden werden Sie dazu eingeladen, an der klausurbegleitenden Befragung teilzunehmen, die Ihre persönliche Einschätzung zu den beiden Frageformaten darstellen soll. Dieser Fragebogen wird anonym sowie unabhängig von Ihrem Klausurergebnis erhoben und ausgewertet.

Bitte bewerten Sie die vorformulierten Aussagen wahrheitsgerecht gemäß Ihrer persönlichen Empfindung. Pro Aussage soll nur ein Kreuz/Markierung auf der fünfstufigen Skala gesetzt werden.

1. **Die Fragen im Single-Best-Answer-Format („Multiple-Choice-Fragen") waren für mich einfacher zu beantworten, als die Fragen im Very-Short-Answer-Format („offen-gestellte-Fragen").**

trifft voll zu	*trifft eher zu*	*trifft teilweise zu*	*trifft kaum zu*	*trifft nicht zu*
O	O	O	O	O

2. **Das Very-Short-Answer-Format bietet eine repräsentativere Darstellung davon, wie ich Inhalte und Fragen in der klinischen Praxis beantworten muss.**

trifft voll zu	*trifft eher zu*	*trifft teilweise zu*	*trifft kaum zu*	*trifft nicht zu*
O	O	O	O	O

3. **Für eine Klausur, die im reinen Very-Short-Answer-Format gestellt ist, würde sich meine Lernstrategie und Vorbereitung ändern.**

trifft voll zu	*trifft eher zu*	*trifft teilweise zu*	*trifft kaum zu*	*trifft nicht zu*
O	O	O	O	O

4. **Ich würde den Gebrauch von Very-Short-Answer-Fragen gegenüber Single-Best-Answer-Fragen in Klausuren bevorzugen.**

trifft voll zu	*trifft eher zu*	*trifft teilweise zu*	*trifft kaum zu*	*trifft nicht zu*
O	O	O	O	O

Vielen Dank für Ihre Teilnahme an dieser Studie!

7 Literaturverzeichnis

Aebli H (Hrsg.): Grundlagen des Lehrens – eine allgemeine Didaktik auf psychologischer Grundlage. 4. Auflage; Klett-Cotta, Stuttgart 1997

Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S (2002): General Competencies And Accreditation In Graduate Medical Education. Health Affairs 21, 103–111

Black P, Wiliam D (1998): Assessment and Classroom Learning. Assess Educ 5, 7–74

Brozo W, Schmelzer R, Spires H: A study of test-wiseness clues in college and university teacher-made tests. Tech Report (84-01) Atlanta, GA 1984

Bundesrat (2017): Verordnung zur Neuregelung der zahnärztlichen Ausbildung; Online verfügbar über: https://www.bundesrat.de/SharedDocs/drucksachen/2017/0501-0600/592-17.pdf?__blob=publicationFile&v=5; abgerufen am 28.07.2020

Centeno A, Primogerio C, Llull L (2007): The process of learning during an examination. Med Educ 41, 619–619

Chenot J, Ehrhardt M (2003): Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: Eine Alternative zur Klausur. Z Allg Med 79, 437–442

Damjanov I, Fenderson BA, Veloski JJ, Rubin E (1995): Testing of medical students with open-ended, uncued questions. Hum Pathol 26, 362–365

Döring N, Bortz J (Hrsg.): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften (Springer-Lehrbuch). 5. Auflage; Springer Berlin Heidelberg, Berlin, Heidelberg 2016, 81-119

Downing SM, Haladyna TM (Hrsg.): Handbook of Test Development. Twelve steps for effective test development. Lawrence Erlbaum Associates Publishers, Mahwah 2006

Eagle M, Leiter E (1964): Recall and recognition in intentional and incidental learning. J Exp Psychol 68, 58–63

Elstein AS (1993): Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. Acad Med 68, 244–9

Entwistle NJ, Meyer JH (1992): Findings and implications from research on student learning. S Afr Med J 81, 593–595

Epstein RM (2007): Assessment in medical education. N Engl J Med 356, 387–396

Farley JK (1989): The multiple-choice test: writing the questions. Nurse Educ 14, 10–12

Feldt LS (1980): A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika 45, 99–105

Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E (1997): The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Hum Pathol 28, 526–532

Fischer MR, Holzer M, Jünger J (2010): Prüfungen an den medizinischen Fakultäten - Qualität, Verantwortung und Perspektiven. GMS Z Med Ausbild 27(5), 703

Gerhard-Szep S, Güntsch A, Pospiech P, Söhnel A, Scheutzel P, Wassmann T, Zahn T (2016): Assessment formats in dental medicine: An overview. GMS J Med Educ 33(4), 1064

Haigh C (2003): An evaluation of a judgmental model of assessment for assessing clinical skills in MSc students. Nurse Educ Pract 3, 43–48

Haladyna, TM (Hrsg.): Developing and Validating Multiple-choice Test Items. 3. Auflage; Lawrence Erlbaum Associates, Mahwah 2004

Hudson JN, Bristow DR (2006): Formative assessment can be fun as well as educational. Adv Physiol Educ 30, 33–37

Krasne S, Wimmers PF, Relan A, Drake TA (2006): Differential Effects of Two Types of Formative Assessment in Predicting Performance of First-year Medical Students. Adv Health Sci Educ Theory Pract 11, 155–171

Leach DC (2002): Competence Is a Habit. JAMA 287, 243

McConnell MM, St-Onge C, Young ME (2015): The benefits of testing for learning on later performance. Adv in Health Sci Educ 20, 305–320

McCoubrie P (2004): Improving the fairness of multiple-choice questions: a literature review. Med Teach 26, 709–712

Millman J, Bishop C, Ebel R (1965): An analysis of test-wiseness. Educ Psychol Meas, 707–726

Möltner A, Schellberg D, Jünger J (2006): Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild 23, 11

Möltner A, Duelli R, Resch F, Schultz J-H, Jünger J (2010): Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten. GMS Z Med Ausbild 27 (3), 703

Moosbrugger H, Kelava A: Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger H, Kelava A (Hrsg.): Testtheorie und Fragebogenkonstruktion (Springer-Lehrbuch). 2. Auflage; Springer Berlin Heidelberg, Berlin, Heidelberg 2012, 7–26

Newble DI, Jaeger K (1983): The effect of assessments and examinations on the learning of medical students. Med Educ 17, 165–171

Newble DI, Entwistle NJ (1986): Learning styles and approaches: implications for medical education. Med Educ 20, 162–175

Pell G, Fuller R, Homer M, Roberts T (2010): How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. Med Teach 32, 802–811

Prüfungsrichtlinie der Universitätsmedizin Göttingen, Präambel Seite 1. Online verfügbar über: https://www.med.uni- goettingen.de/de/media/G1-2_lehre/studium_ studienordnung_ pruefungsrichtlinie.pdf; abgerufen am 03.08.2020

Raupach T, Hanneforth N, Anders S, Pukrop T, Th J ten Cate O, Harendza S (2010): Impact of teaching and assessment format on electrocardiogram interpretation skills: Electrocardiogram training and assessment. Med Educ 44, 731–740

Raupach T, Brown J, Anders S, Hasenfuss G, Harendza S (2013): Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. BMC Med 11, 61

Rogers WT, Yang P (1996): Test-Wiseness: Its Nature and Application. Eur J Psychol Assess 12, 247–259

Sadler DR (1989): Formative assessment and the design of instructional systems. Instr Sci 18, 119–144

Sam AH, Hameed S, Harris J, Meeran K (2016): Validity of very short answer versus single best answer questions for undergraduate assessment. BMC Med Educ 16, 266

Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, Harris J, Meeran K (2018): Very-short-answer questions: reliability, discrimination and acceptability. Med Educ 52, 447–455

Sam AH, Westacott R, Gurnell M, Wilson R, Meeran K, Brown C (2019a): Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. BMJ Open 9, e032550

Sam AH, Fung CY, Wilson RK, Peleva E, Kluth DC, Lupton M, Owen DR, Melville CR, Meeran K (2019b): Using prescribing very short answer questions to identify sources of medication errors: a prospective study in two UK medical schools. BMJ Open 9, e028863

Schuwirth LWT, Vleuten CPM, Donkers HHLM (1996): A closer look at cueing effects in multiple-choice questions. Med Educ 30, 44–49

Shumway JM, Harden RM, Association for Medical Education in Europe (2003): AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach 25, 569–584

Sopka S, Simon M, Beckers SK: »Assessment drives Learning«: Konzepte zur Erfolgs- und Qualitätskontrolle. In: St.Pierre M, Breuer G (Hrsg.): Simulation in der Medizin. Springer Berlin Heidelberg, Berlin, Heidelberg 2013, 83–92

Thoma G-B, Köller O (2018): Test-wiseness: ein unterschätztes Konstrukt?: Empirische Befunde zur Überprüfung und Erlernbarkeit von test-wiseness. Z f Bildungsforsch 8, 63–80

UMG Lehre, Studiengang Zahnmedizin Aufbau. Online verfügbar über: https://www.umg.eu/studium-lehre/studiengaenge/zahnmedizin/; abgerufen am 08.09.2020

van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, Vos J (2012): Preferred question types for computer-based assessment of clinical reasoning: a literature study. Perspect Med Educ 1, 162–171

Van Der Vleuten CPM (1996): The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1, 41–67

Veloski JJ, Rabinowitz HK, Robeson MR, Young PR (1999): Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. Acad Med 74, 539–546

Wass V, Van der Vleuten C, Shatzer J, Jones R (2001): Assessment of clinical competence. The Lancet 357, 945–949

Wood T (2009): Assessment not only drives learning, it may also help learning. Med Educ 43, 5–6

Danksagung

Ich möchte mich insbesondere bei Herrn Prof. Dr. med. Tobias Raupach für die zuverlässige wissenschaftliche Betreuung, die Überlassung des Themas und die Unterstützung in dieser Promotionsarbeit bedanken. Herrn Prof. Dr. med. dent. Ralf Bürgers danke ich für die Ermöglichung der Promotionsarbeit.

Ein besonderer Dank gilt auch Herrn Dr. med. dent. Torsten Wassmann für die stetige Unterstützung und engagierte Zusammenarbeit.