Florian Klapproth[1]

Welche Rolle spielen Distraktoren in Multiple-Choice-Aufgaben für die Testqualität?

Zusammenfassung

Diese Studie untersucht den Einfluss von Distraktoren in Multiple-Choice-Aufgaben auf die Testqualität. Anhand einer Stichprobe von $N = 128$ Studierenden wurden drei Versionen eines Tests verglichen, bei denen die Antwort „Keine der genannten Optionen" entweder als Distraktor, als korrekte Antwort oder gar nicht eingesetzt wurde. Die Ergebnisse zeigen, dass die Anwesenheit dieser Antwort die Testschwierigkeit erhöht und die Distraktoreffizienz steigert, ohne die Trennschärfe oder Reliabilität signifikant zu beeinflussen. Diese Ergebnisse deuten darauf hin, dass die Antwort „Keine der genannten Optionen" die Testqualität zumindest nicht verschlechtert, solange sie mit Bedacht eingesetzt wird.

Schlüsselwörter

Multiple-Choice-Aufgaben, Distraktor, „Keine der genannten Optionen", Itemschwierigkeit, Testqualität

1 Medical School Berlin; florian.klapproth@medicalschool-berlin.de; ORCiD 0000-0002-4598-837X

https://doi.org/10.21240/zfhe/20-1/12

Which role play distractors in multiple-choice items for the quality of the test?

Abstract

This study examines the influence of distractors in multiple-choice tasks on test quality. Using a sample of $N = 128$ students, three versions of a test were compared in which the answer "None of the above" was used either as a distractor, as a correct answer, or not at all. The results show that the presence of this answer increases test difficulty and distractor efficiency without significantly affecting selectivity or reliability. These results indicate that the option "None of the above" does not worsen test quality, at least as long as it is used with caution.

Keywords

multiple-choice items, distractor, "None of the above", item difficulty, test quality

1 Einleitung

Prüfungen spielen eine zentrale Rolle in der Ausbildung von Studierenden, da sie sowohl Lehrenden als auch Lernenden Rückmeldung über den Leistungsstand geben und dadurch mögliche Verbesserungsmaßnahmen aufzeigen (Linder et al., 2018). Insbesondere schriftliche Prüfungen bieten zahlreiche ökonomische Vorteile und sind daher gegenüber mündlichen Prüfungen attraktiver geworden. Durch die Notwendigkeit einer Zertifizierung jedes Leistungsmoduls hat sich das Prüfungsaufkommen an Hochschulen in den letzten Jahren deutlich erhöht (Lindner et al., 2018). Die vorliegende Studie soll einen Beitrag zur Verbesserung der Qualität schriftlicher Prüfungen liefern.

Um den steigenden Anforderungen gerecht zu werden, sind Testentwickler:innen auch an ökonomischen Vorteilen interessiert, um die wachsende Zahl der Testteilnehmer:innen bewältigen zu können und den Aufwand für die Testauswertung zu reduzieren. In diesem Zusammenhang erfreuen sich Multiple-Choice-Tests zunehmender Beliebtheit. Sie bieten den Vorteil einer schnellen Leistungsbewertung sowie einer erhöhten Objektivität und Zuverlässigkeit (DiBattista et al., 2014). Durch die Umstrukturierung der Studiengänge und die steigende Anzahl der Prüfungen haben sich geschlossene Aufgabenformate, wie Multiple-Choice-Tests, gegenüber offenen Formaten oftmals durchgesetzt (Xu et al., 2016). Diese Entwicklung stößt jedoch nicht überall auf Zustimmung. Geschlossenen Aufgabenformaten (wie Multiple-Choice-Tests) wird zum einen unterstellt, sie seien nicht in der Lage, Kreativität und praktische Fähigkeiten valide zu erfassen (Brauns & Schubert, 2008), und zum anderen, dass sie oberflächliches Lernen begünstigten (Knott, 2008).

Trotz ihrer ökonomischen Vorteile bringt die Verwendung von Multiple-Choice-Tests auch Herausforderungen mit sich. Eine der größten Herausforderungen besteht in der Erstellung qualitativ hochwertiger Fragen, insbesondere wenn begrenzte zeitliche Ressourcen zur Verfügung stehen. Dies kann dazu führen, dass Multiple-Choice-Aufgaben ihr angestrebtes Qualitätsniveau nicht vollständig erreichen und hinter den Erwartungen zurückbleiben (Linder et al., 2018). Daher ist es von großer

Bedeutung, ein einheitliches und hohes Qualitätsniveau sicherzustellen, um eine valide Diagnostik und verlässliche Aussagen über die Prüfungsleistungen von Studierenden treffen zu können.

Die Qualität von Multiple-Choice-Aufgaben hängt unter anderem von den Merkmalen der Distraktoren ab. Distraktoren sind die inkorrekten Antwortoptionen. Sie sollen es unwissenden Prüflingen erschweren, zufällig die korrekte Antwort zu wählen. Es besteht eine komplexe Beziehung zwischen korrekten Optionen und Distraktoren. Diese rührt daher, dass Testpersonen die richtige Antwort in der Regel durch die wechselseitige Abgrenzung der einzelnen Antwortoptionen wählen (Klapproth, 2023). So konnten Little und Bjork (2015) zeigen, dass Distraktoren den Erinnerungsprozess und damit die Wahl der korrekten Antwort begünstigen, wenn sie plausibel sind und in einigen Aspekten der korrekten Option inhaltlich ähneln. Distraktoren können den Erinnerungsprozess allerdings auch erschweren, nämlich dann, wenn sie keinerlei Zusammenhang mit der korrekten Antwort aufweisen (Bishara & Lanzo, 2015).

Kontroverse Diskussionen bestehen über die Verwendung von summarischen Optionen (vgl. Gierl et al., 2017; Haladyna & Rodriguez, 2013). Mit summarischen Optionen werden mehrere Optionen in einer zusammengefasst. Dies erfolgt in der Regel über die Option „Alle der oben genannten Optionen“ (engl. All of the Above [AOTA]) oder „Keine der oben genannten Optionen“ (engl. None of the Above [NOTA]). Wenn eine Testperson NOTA auswählt, drückt sie damit aus, dass ihrer Meinung nach keine der angebotenen Antwortmöglichkeiten korrekt ist. Ein Grund für die Verwendung von NOTA liegt unter anderem in der Absicht, die Testschwierigkeit zu erhöhen, um somit besser die Fähigkeit einer Testperson erfassen zu können (García-Pérez, 1993).

Ein Argument gegen die Verwendung von summarischen Optionen besteht darin, dass sie häufig von Tertpersonen gewählt werden, die nur über partielles Wissen verfügen und die korrekte Antwort nicht kennen. Insbesondere gegenüber der Verwendung von NOTA existieren ernstzunehmende Vorbehalte (Gross, 1994). Wenn in einer Multiple-Choice-Frage die korrekte Antwort NOTA ist und die Testperson

die Antwort auf die Frage nicht weiß, so ist sie geneigt, die summarische Option zu wählen und wird dafür mit einem Punkt belohnt.

Die empirische Befundlage zum Einfluss von NOTA auf die Qualität von Multiple-Choice-Aufgaben ist allerdings gemischt. Relativ konsistent sind die Ergebnisse in Bezug auf die Schwierigkeit von Multiple-Choice-Items mit NOTA als Option, wenn die NOTA-Option die korrekte Antwort darstellt. Die Schwierigkeit ist einer Reihe von Studien zufolge höher im Vergleich zu Multiple-Choice-Items ohne diese Option (Sanderson, 2010). Die erhöhte Schwierigkeit bei Items, in denen NOTA die korrekte Option ist, lässt sich dadurch erklären, dass die korrekte Antwort in diesen Items nicht dargeboten wird und somit das Item statt einer Wiedererkennungsaufgabe (Rekognition) nun eine Aufgabe des freien Erinnerns (Recall) darstellt (DiBattista et al., 2014). Rekognitionsaufgaben sind im Allgemeinen leichter als Recall-Aufgaben (Funk & Dickson, 2011). Ein Grund für die generell erhöhte Schwierigkeit von Multiple-Choice-Items mit NOTA als Option könnte darin liegen, dass Testpersonen die inhaltlich bestimmten Optionen untereinander vergleichen und bei Nichtwissen diejenige wählen, die der korrekten Antwort inhaltlich am ähnlichsten zu sein scheint (DiBattista et al., 2014). Demnach sollte die korrekte Option seltener gewählt werden (und damit das Item schwerer sein), wenn NOTA die korrekte Option darstellt, als wenn NOTA ein Distraktor ist.

Unklarer scheint der Effekt von NOTA auf die Schwierigkeit zu sein, wenn die NOTA-Option ein Distraktor ist. Während einige Studien gezeigt haben, dass NOTA auch dann die Schwierigkeit erhöht, wenn es eine inkorrekte Option darstellt (Frary, 1991), finden sich in anderen Studien Hinweise darauf, dass NOTA als inkorrekte Option keinen Effekt auf die Schwierigkeit hat (Little, 2023; Pachai et al., 2015).

Neben der Schwierigkeit ist seine Trennschärfe eine Schlüsseldeterminante für die Qualität eines Multiple-Choice-Items (Ebel, 1975). Sie spiegelt das Ausmaß wider, in welchem leistungsstarke Testpersonen mit größerer Wahrscheinlichkeit als leistungsschwache Testpersonen die korrekte Antwort auswählen. Von einem trennscharfen Item wird erwartet, dass „fähige“ Testpersonen das Item lösen, während „unfähige“ Testpersonen das Item nicht lösen. Bezüglich der Trennschärfe ist die

Befundlage zu NOTA weniger eindeutig. Einige Studien zeigten keinen Unterschied zwischen Items mit und ohne NOTA als Option (Crehan & Haladyna, 1991; Sanderson, 2010), andere fanden, dass Items mit NOTA weniger trennscharf sind (Wesman & Bennett, 1946), wiederum andere zeigten den gegenteiligen Effekt (Rich & Johanson, 1990). Ein Argument für eine geringere Trennschärfe von Items, die NOTA als Option beinhalten, könnte darin bestehen, dass eine Erhöhung der Schwierigkeit tendenziell mit einer Verringerung der Trennschärfe einhergeht, weil (sehr) schwere Items schlechter als mittelschwere Items zwischen „fähigen" und „unfähigen" Testpersonen diskriminieren können.

Auch die Reliabilität eines Tests, also seine Messgenauigkeit, könnte prinzipiell durch die Verwendung von Items mit der Option NOTA beeinflusst werden. Wenn Items durch die Anwesenheit von NOTA häufiger falsch beantwortet werden (und somit die Schwierigkeit steigt und die Trennschärfe sich verringert), dann sollte auch die Reliabilität des Tests als Ganzem geringer werden, da die Beantwortung der Items stärker auf den Zufall als auf Wissen zurückgeht. Belastbare Evidenzen dazu sind allerdings kaum vorhanden, nur wenige Studien haben sich mit dieser Frage befasst. Die vorliegenden Daten sprechen eher dafür, dass die Testreliabilität weitgehend unbeeinflusst von NOTA ist (z. B. Atalmış & Kingston, 2017; Rich & Johanson, 1990).

Distraktoren können unterschiedlich effizient sein. Distraktoreffizienz meint dabei, mit welcher Häufigkeit Testpersonen Distraktoren in einem Multiple-Choice-Item wählen. Distraktoren, die weniger als in 5 % aller Darbietungen gewählt werden, gelten als ineffizient (Tarrant et al., 2009). Ziel einer Testentwicklung sollte sein, möglichst effiziente Distraktoren zu verwenden, da durch diese sichergestellt werden kann, dass „unfähige" Testpersonen die korrekte Option seltener wählen als eine der inkorrekten Optionen und der Test dadurch besser diskriminiert. Inwieweit die Verwendung von NOTA als Option die Distraktoreffizienz beeinflusst, ist bislang kaum untersucht worden. Anzunehmen ist, dass eine höhere Schwierigkeit auch mit einer höheren Distraktoreffizienz einhergeht, da in schwierigen Items die korrekte Antwort seltener und daher die Distraktoren häufiger gewählt werden als in leichten Items.

Das Ziel der vorliegenden Arbeit bestand darin, den Einfluss der Antwortoption NOTA auf die Testqualität zu prüfen. Diese wurde operationalisiert als (1) Testschwierigkeit, (2) die über alle Items gemittelte Trennschärfe, (3) Testreliabilität und (4) Distraktoreffizienz eines Multiple-Choice-Tests. Mit dieser Differenzierung der Testqualität unterscheidet sich die vorliegende Studie von bisherigen Studien, die den Einfluss von NOTA auf die Testqualität untersucht haben. Darüber hinaus wurde in der vorliegenden Arbeit zwischen der Funktion von NOTA als Distraktor und NOTA als korrekte Antwort unterschieden. Schließlich liegt mit der vorliegenden Studie eine der wenigen Arbeiten überhaupt vor, die im deutschsprachigen Raum zur Analyse von Multiple-Choice-Aufgaben durchgeführt wurden.

In dieser Studie wurden drei Versionen von Multiple-Choice-Tests verglichen: Version A enthielt nur inhaltliche Optionen und kein NOTA als Option und diente damit als Kontrollversion. In Version B war NOTA ein Distraktor, während in Version C NOTA die korrekte Option darstellte.

Es wurde angenommen, dass alle vier Indizes der Testqualität miteinander zusammenhängen. Zunächst wurde erwartet, dass die Option NOTA im Vergleich zu inhaltlich bestimmten Optionen den Test erschwert (Frary, 1991). Die Erschwernis sollte dann am größten sein, wenn NOTA die korrekte Option darstellt (Pachai et al., 2015). Da eine hohe Testschwierigkeit mit einer eher geringen Trennschärfe einhergeht, wurde außerdem erwartet, dass die Trennschärfe am höchsten in der Testversion ohne NOTA und am geringsten in der Testversion mit NOTA als korrekter Option ist. Da auch zwischen Reliabilität und Trennschärfe eine positive Beziehung besteht (Yousfi, 2005), wurde erwartet, dass die Reliabilität von Testversion A am höchsten und die von Testversion C am geringsten ist. Schließlich wurde in Bezug auf die Distraktoreffzienz angenommen, dass diese mit erhöhter Schwierigkeit ebenfalls höher wird, so dass sie in Testversion A am geringsten und in Testversion C am höchsten ausfallen sollte.

2 Methode

2.1 Versuchspersonen

Mit Hilfe des Programms g*power (Faul et al., 2009) wurde bei Erwartung eines mittleren Effekts ($d = .50$), einer Alpha-Fehlerwahrscheinlichkeit von $\alpha = .05$ und einer Teststärke von $1\text{-}\beta = .80$ ein erforderlicher Stichprobenumfang von $N = 153$ ermittelt. Insgesamt nahmen $N = 128$ Versuchspersonen an der Untersuchung teil. Das mittlere Alter der Versuchspersonen betrug 23.6 Jahre ($SD = 3.5$). Die Versuchspersonen wurden zufällig drei Bedingungen (Version A, B, C) zugewiesen. Alle Versuchspersonen waren zum Zeitpunkt der Untersuchung an Universitäten eingeschriebene Studierende. 75.8 % der Versuchspersonen studierten Psychologie, 10.9 % Wirtschaftswissenschaften, und die verbleibenden 13.3 % verteilten sich auf naturwissenschaftliche, sozialwissenschaftliche und andere Studiengänge. Alle Versuchspersonen gaben an, bereits Erfahrungen mit Multiple-Choice-Aufgaben gesammelt zu haben. Für die Durchführung der Untersuchung wurde bei der hochschulinternen Ethikkommission ein Ethikvotum beantragt, welches am 08.05.2024 unter der Kennung MSB-2024/171 erteilt wurde.

2.2 Versuchsmaterial

Die verwendeten Items stellten eine Auswahl aus dem Hohenheimer Inventar zum Politikwissen (Trepte et al., 2017) dar. Das Inventar enthält 85 Items und erfasst allgemeines politisches Wissen aus unterschiedlichen inhaltlich-historischen Kategorien. Die aus diesem Inventar für die vorliegende Studie ausgewählten 30 Items prüften politisches und historisches Wissen über Ereignisse und Fakten vom 18. Jahrhundert bis heute. Jedes Item enthielt vier Optionen, von denen immer nur eine richtig war.

Jedes der 30 Original-Items wurde für die drei Versionen angepasst. In Version A blieb das Item unverändert; allerdings wurde die Reihenfolge der Optionen zufällig variiert. Für Version B wurde ein Distraktor des Original-Items zufällig ausgewählt

und durch NOTA ersetzt. Für Version C wurde die korrekte Option durch NOTA ersetzt.

Tabelle 1 zeigt ein Beispiel-Item für die drei Versionen.

Item-Stamm (Original)	**Version A (kein NOTA)**	**Version B (NOTA als Distraktor)**	**Version C (NOTA als korrekte Option)**
Nach dem Zweiten Weltkrieg war Deutschland ein geteiltes Land und es gab zwei deutsche Staaten. Wie war ihre offizielle Bezeichnung?	**a) Bundesrepublik Deutschland und Deutsche Demokratische Republik** b) Ost- und Westdeutschland c) Bundesdeutsche Demokratie und Deutsche kommunistische Republik d) Ostdeutsche Republik und Westdeutsche Republik	**a) Bundesrepublik Deutschland und Deutsche Demokratische Republik** b) Ost- und Westdeutschland c) Bundesdeutsche Demokratie und Deutsche kommunistische Republik d) keine der oben genannten Optionen	a) Ost- und Westdeutschland b) Bundesdeutsche Demokratie und Deutsche kommunistische Republik c) Ostdeutsche Republik und Westdeutsche Republik **d) keine der oben genannten Optionen**

Tabelle 1: Beispiel-Item in den drei Versionen
Anmerkung: Die korrekte Option ist fettgedruckt.

Die Versionen der drei Bedingungen unterschieden sich darin, dass die NOTA-Option nur in den Versionen B und C vorkam, während in Version A an ihrer Stelle eine inhaltlich bestimmte Option stand. In Version B war bei den Items 1 bis 10 NOTA die korrekte Option und bei den Items 11 bis 30 NOTA ein Distraktor. In Version C war dagegen bei den Items 1 bis 10 NOTA ein Distraktor und bei den Items 11 bis 30 NOTA die korrekte Option. Durch die Variation der Funktion von NOTA (Distraktor oder korrekte Antwort) sollte sichergestellt werden, dass die Versuchspersonen kein Antwortmuster erkennen und so beispielsweise in Version C entdecken, dass NOTA immer die korrekte Option darstellt.

2.3 Versuchsdurchführung

Die Untersuchung wurde über SoSci Survey, einer Onlineplattform für die Erstellung von Umfragen, durchgeführt. Die Rekrutierung der Versuchspersonen erfolgte über die Verbreitung des Teilnahmelinks über die Plattform Instagram sowie über den Instant-Messenger-Dienst WhatsApp. Die Daten wurden im Zeitraum vom 09.05.2024 bis zum 05.07.2024 erhoben. Alle Teilnehmenden führten die Untersuchung selbstständig, mit ihren eigenen elektronischen Endgeräten und ohne Aufsicht durch. Die Teilnahme war freiwillig, anonym und ohne finanzielle Entlohnung.

Alle Teilnehmenden wurden zu Beginn der Untersuchung auf einer Startseite über den Untersuchungszweck, die Durchführungsdauer, Freiwilligkeit und Anonymität aufgeklärt sowie über die Verwendung ihrer Daten informiert. Die Teilnahmeinformationen mussten von allen Versuchspersonen vor Beginn der Untersuchung mit einer Einwilligungserklärung bestätigt werden. Zudem erhielten alle Teilnehmenden die Information, dass sie die Untersuchung jederzeit und ohne Nachteile abbrechen können.

Nach der Studieninformation wurden die Versuchspersonen zufällig den drei Bedingungen zugeteilt. Anschließend wurden die Items aus dem Hohenheimer Inventar zum Politikwissen dargeboten. Die Reihenfolge der Itemdarbietung war zufällig. Den Abschluss der Studie bildeten die Erfassung der soziodemografischen Daten und eine Danksagung.

2.4 Analysen

Für jede Testversion (A, B, C) wurden auf Itemebene eine Distraktoranalyse, die Analyse der Schwierigkeit und die Trennschärfenanalyse sowie auf Testebene eine Reliabilitätsanalyse durchgeführt.

Für die Distraktoranalyse wurden sogenannte nonfunktionale Distraktoren ermittelt. Darunter werden Distraktoren verstanden, die von weniger als 5 % der Versuchspersonen als Antwort gewählt wurden und somit als wenig plausibel bzw. wenig diskri-

minierend gelten können (Tarrant et al., 2009). Die Anzahl nonfunktionaler Distraktoren innerhalb eines Items ergab die Distraktoreffizienz (DE). Wenn ein Item keinen nonfunktionalen Distraktor beinhaltete, war die DE maximal und damit gleich 1. Mit jedem nonfunktionalen Distraktor verringerte sich die DE um 1/n, wobei n gleich der Anzahl der Antwortoptionen war. In der vorliegenden Studie entsprach ein nonfunktionaler Distraktor daher einer DE von 1 – 1/3 = 2/3, zwei nonfunktionale Distraktoren einer DE von 1 – 2/3 = 1/3 und drei nonfunktionale Distraktoren einer DE von 1 – 3/3 = 0.

Für die Ermittlung der Itemschwierigkeit wurde der Schwierigkeitsindex *P* verwendet. Dieser gibt an, wie groß der Anteil von Personen ist, die ein Item richtig beantwortet haben. Ein hoher Wert spiegelt daher ein leichtes Item wider.

Für die Bestimmung der Trennschärfe wurde der Trennschärfeindex verwendet. Er gibt an, inwieweit ein einzelnes Item in der Lage ist, das Gesamttestergebnis vorherzusagen. Die operationale Definition für den Trennschärfeindex ist die Korrelation zwischen der Antwort auf ein Item und dem vom Itemwert bereinigten Gesamttestwert. Für die Analyse der Testreliabilität wurde auf McDonalds Omega als Maß der internen Konsistenz zurückgegriffen.

3 Ergebnisse

In Tabelle 2 werden die mittleren Testscores sowie die daraus berechneten mittleren Item-Schwierigkeiten für die einzelnen Bedingungen berichtet.

	Testscore		Itemschwierigkeit *P*		
Bedingung	***M***	***SD***	***M***	***SD***	***n***
Version A (Kein NOTA)	22.51	5.74	.75	.19	43
Version B (NOTA als Distraktor)	13.23	6.52	.44	.22	40
Version C (NOTA als korrekte Option)	14.73	6.30	.49	.21	45

Tabelle 2: Testscores und Itemschwierigkeiten für die drei experimentellen Bedingungen

In der Bedingung ohne die Option NOTA wurden deutlich mehr Punkte erzielt, also deutlich häufiger korrekt geantwortet, als es in den Bedingungen mit NOTA der Fall war. Entsprechend war die mittlere Itemschwierigkeit in Bedingung A geringer, der *P*-Index folglich höher, als in den Bedingungen B und C. Der globale *F*-Test bestätigt diesen Effekt, $F(2, 125) = 27.47$, $p < .001$, $\eta^2 = .31$. In geplanten Kontrasten zeigte sich, dass der Unterschied zwischen Bedingung A und den Bedingungen B und C signifikant war, $t(125) = 7.36$, $p < .001$, $d = 1.38$. Der Unterschied zwischen Bedingung B und Bedingung C war hingegen nicht signifikant, $t(125) = -1.12$, $p = .264$, $d = -.12$.

Für die Prüfung auf Unterschiede zwischen den Trennschärfen und McDonalds Omega wurden beide Kennwerte für das Hypothesentesten zunächst Fisher-*Z*-transformiert. Anschließend wurden die Unterschiede zwischen den Bedingungen mittels *Z*-Test einseitig geprüft. Tabelle 3 zeigt die über alle Items gemittelten Trennschärfen und McDonalds Omega für die drei Bedingungen.

Bedingung	Trennschärfe	McDonalds Omega	*n*
Version A (Kein NOTA)	.449	.904	43
Version B (NOTA als Distraktor)	.402	.879	40
Version C (NOTA als korrekte Option)	.378	.865	45

Tabelle 3: Trennschärfe und McDonalds Omega

Wie in Tabelle 3 deutlich wird, wurden in Bedingung A (kein NOTA) die höchste mittlere Trennschärfe und die höchste Reliabilität erzielt. Allerdings war der Unterschied hinsichtlich der Trennschärfe zwischen Bedingung A und Bedingung B, $Z = 0.25$, $p = .401$, zwischen Bedingung A und Bedingung C, $Z = 0.39$, $p = .349$, sowie zwischen Bedingung B und Bedingung C, $Z = 0.13$, $p = .450$, nicht signifikant.

Ähnliches zeigte sich hinsichtlich der Reliabilität, da der Unterschied zwischen Bedingung A und B, $Z = 0.54$, $p = .296$, zwischen Bedingung A und C, $Z = 0.82$, $p = .207$, sowie zwischen Bedingung B und C, $Z = 0.26$, $p = .396$, ebenfalls nicht signifikant war.

Tabelle 4 zeigt die Ergebnisse der Distraktoranalyse. Augenscheinlich waren die Distraktoren im Durchschnitt in den beiden NOTA-Bedingungen deutlich effizienter als in der Bedingung ohne NOTA. Während in Version A durchschnittlich etwa zwei Drittel aller Antworten auf die korrekte Option fielen, waren es in Version B nur etwas mehr als ein Drittel und in Version B immer noch deutlich weniger als die Hälfte. Interessanterweise fielen in Version B mehr als die Hälfte der Falschantworten auf die Option NOTA.

Bedingung	% korrekt	% Distraktor	Mittlere Distraktor-effizienz	*n*
Version A (Kein NOTA)	68.7	31.3	0.45	43
Version B (NOTA als Distraktor)	37.5	62.5 (33.6 NOTA)	0.90	40
Version C (NOTA als korrekte Option)	42.7	57.3	0.90	45

Tabelle 4: Ergebnisse der Distraktoranalyse

Da die Distraktoreffizienz ähnlich wie ein Korrelationskoeffizient nur Werte zwischen 0 und 1 annehmen kann, wurde auch hier der Unterschied zwischen den Bedingungen mit dem Fisher-*Z*-Test geprüft. Es zeigte sich, dass der Unterschied zwischen Bedingung A und Bedingung B, $Z = -4.36$, $p < .001$, sowie zwischen Bedingung A und Bedingung C, $Z = -4.52$, $p < .001$, statistisch signifikant war, der Unterschied zwischen Bedingung B und C jedoch nicht, $Z = -0.02$, $p = .981$.

4 Diskussion

Ziel der vorliegenden Studie war es, den Einfluss der Option NOTA auf die Testqualität zu untersuchen. Die gewonnenen Ergebnisse zeigen deutlich, dass NOTA die Testqualität beeinflusst. Das augenfälligste und die vorausgehende Hypothese bestätigende Ergebnis war der Effekt von NOTA auf die Testschwierigkeit: Die Anwesenheit von NOTA als Option erschwerte den Test insgesamt. Entgegen der Hypothese zeigte sich jedoch kein Unterschied in der Testschwierigkeit zwischen den beiden NOTA-Bedingungen – es spielte also keine Rolle, ob NOTA die korrekte oder eine inkorrekte Option war. Dass Items generell schwieriger werden, wenn

NOTA die korrekte Option ist, lässt sich dadurch erklären, dass die Aufgabe von einer Rekognitions- zu einer Recall-Aufgabe wird, da die korrekte Antwort nicht dargeboten wird und somit nicht wiedererkannt werden kann. Eine mögliche Erklärung dafür, dass Items auch dann schwieriger werden, wenn NOTA lediglich ein Distraktor ist, könnte darin bestehen, dass die Anwesenheit von NOTA unabhängig von ihrer Funktion dazu führt, dass Testpersonen, die unsicher über die richtige Antwort sind, die vorhandenen Optionen sorgfältiger bewerten und weniger geneigt sind, spontan die plausibelste Antwort zu wählen (Frary, 1991; Rodriguez, 1997).

Hypothesenkonform war auch der Effekt von NOTA auf die Trennschärfe: diese verringerte sich bei Anwesenheit von NOTA als Option. Allerdings war der Unterschied nicht statistisch signifikant. Gleiches zeigte sich auch in Bezug auf die Testreliabilität, die am höchsten in der Bedingung ohne NOTA war. Allerdings waren auch hier die Unterschiede zwischen den Bedingungen nicht signifikant. Das Ausbleiben eines signifikanten Unterschieds im Hinblick auf Trennschärfe und Reliabilität ist möglicherweise auch dem Umstand geschuldet, dass beide Maße Korrelationen sind und der Vergleich von Korrelationen große Stichproben erfordert, um vorhandene Effekte aufzudecken (Papenberg & Musch, 2017). Aus diesem Grund erscheint die Teststärke der vorliegenden Studie für die Hypothesenprüfung nicht groß genug gewesen zu sein.

Die erhöhte Schwierigkeit in den Testversionen mit NOTA spiegelte sich auch in einer höheren Distraktoreffizienz wider. Distraktoren statt der korrekten Option wurden deutlich häufiger gewählt, wenn NOTA eine Option war. War NOTA die korrekte Antwort, wurden häufiger Distraktoren gewählt im Vergleich zur Kontrollversion ohne NOTA. War NOTA ein Distraktor, wurde dieser häufiger gewählt als die anderen Distraktoren und im Durchschnitt fast so häufig wie die korrekte Option.

5 Limitationen und Schlussfolgerungen

Eine Limitation dieser Arbeit ist der Stichprobenumfang, der für eine Prüfung von Unterschieden in der Höhe von Trennschärfe und Reliabilität möglicherweise nicht groß genug war (Feldt, 1969), was auch durch den nach der Poweranalyse ermittelten höheren als realisierten Stichprobenumfang nahegelegt wird. In zukünftigen Studien sollten für diesen Zweck größere Stichprobenumfänge realisiert werden. Allerdings deuten die Ergebnisse dieser Studie bereits darauf hin, dass der Effekt, also der Unterschied zwischen den hier realisierten Bedingungen, eher klein ist und somit auch in einer größeren Stichprobe ein eher kleiner Effekt zu erwarten ist.

Eine weitere Limitation besteht darin, dass in der Studie ausschließlich Fragen zum Politikwissen ausgewertet wurden. Inwieweit Generalisierungen auf andere Inhaltsbereiche möglich sind, sollte durch weitere Untersuchungen geprüft werden.

Des Weiteren könnte das Vorwissen der Versuchspersonen in Bezug auf das Politikwissen eine mögliche Störvariable gewesen sein, für die hätte kontrolliert werden müssen. So lässt sich nicht ausschließen, dass Unterschiede in der Testschwierigkeit auch auf Unterschiede im Vorwissen zurückgingen.

NOTA als Option in Multiple-Choice-Aufgaben kann eine Möglichkeit darstellen, die Schwierigkeit der Aufgaben zu erhöhen, ohne dabei die Diskriminationsfähigkeit der Items oder die Reliabilität des Tests wesentlich zu reduzieren. Diesen Schluss legen die Ergebnisse dieser Untersuchung nahe. Eine Erhöhung der Schwierigkeit kann dann sinnvoll sein, wenn das Anforderungsniveau der Prüfung insgesamt angehoben werden soll. Die Beantwortung von Multiple-Choice-Aufgaben, in denen NOTA eine Option ist, erfordert größere kognitive Anstrengungen (Rodriguez, 1997), und das Erkennen von NOTA als korrekte Option lässt darauf schließen, dass Testpersonen die richtige Antwort nicht wiedererkannt, sondern tatsächlich verstanden haben. Gleichwohl sollte die Option NOTA mit Bedacht gewählt werden, da sie nicht immer Sinn ergibt. Das ist beispielsweise der Fall, wenn alle inhaltlichen Optionen die Menge aller logischen Optionen vollständig beinhalten. So würde für die

Frage „Wo liegt in empirischen Studien häufig das Signifikanzniveau?“ mit den folgenden inhaltlichen Optionen „< 5 %“, „5 %“ und „> 5 %“ eine weitere Option keinen Sinn ergeben. Aus praktischer Sicht bieten die Ergebnisse der vorliegende Studien Grund für vorsichtigen Optimismus, was den Einsatz summarischer Optionen wie NOTA betrifft. NOTA scheint vor allem dann eine gute Wahl zu sein, wenn das Generieren inhaltlich bestimmter Distraktoren schwerfällt.

Literaturverzeichnis

Atalmış, E., & Kingston, N. (2017). Three, four, and none of the above options in multiple-choice items. *Turkish Journal of Education*, *6*, 143–157. https://doi.org/10.19128/turje.333687

Bishara, A. J., & Lanzo, L. A. (2015). All of the above: When multiple correct response options enhance the testing effect. *Memory*, *23*, 1013–1028. https://doi.org/10.1080/09658211.2014.946425

Brauns, K., & Schubert, S. (2008). Qualitätssicherung von Multiple-Choice-Prüfungen. *Blickpunkt Hochschuldidaktik, 118*, 92–102.

Crehan, K. D., & Haladyna, T. M. (1991). The validity of two item-writing rules. *The Journal of Experimental Education, 59*, 183–192. https://doi.org/10.1080/00220973.1991.10806560

DiBattista, D., Sinnige-Egger, J., & Fortuna, G. (2014). The “none of the above” option in multiple-choice testing: An experimental study. *The Journal of Experimental Education, 82*, 168–183. https://doi.org/10.1080/00220973.2013.795127

Ebel, R. L. (1975). Can teachers write good true-false items?. *Journal of Educational Measurement, 12*, 31–35.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feldt, L. S. (1969). A test of the hypothesis that cronbach’s alpha or kuder-richardson coefficent twenty is the same for two tests. *Psychometrika, 34*, 363–373.

Frary, Robert B. 1991. The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4*, 115–124.
https://doi.org/10.1207/s15324818ame0402_2

Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology, 38*, 273–277.
https://doi.org/10.1177/0098628311421329

Garcia-Perez, M. A. (1993). In defense of "none of the above". *British Journal of Mathematical and Statistical Psychology, 46, 213–229*.
https://doi.org/10.1111/j.2044-8317.1993.tb01013.x

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors in multiple-choice tests: A comprehensive review. *Review of Educational Research, 87,* 1082–1116. https://doi.org/10.3102/0034654317726529

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: the case of "none of the above." *Evaluation & The Health Professions, 17*, 123–126.
https://doi.org/10.1177/016327879401700108

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hingorjo M. R., & Jaleel F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distracter efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*, 142–147.

Klapproth, F. (2023). *Von den Lehrzielen zur schriftlichen Prüfung. Ein Leitfaden für Lehrende der Psychologie*. Hogrefe. https://doi.org/10.1026/03191-000

Knott, J. (2008). Wie lernen Studierende eigentlich für Multiple-Choice-Klausuren?. *Zeitschrift für Beratung und Studium, 2/2008*, 47–50.

Lindner, M. A., Mayntz, S. M., & Schult, J. (2018). Studentische Bewertungen und Präferenz von Hochschulprüfungen mit Aufgaben im offenen und geschlossenen Antwortformat. *Zeitschrift für Pädagogische Psychologie, 32*, 239–248.
https://doi.org/10.1024/1010-0652/a000229

Little, J. L. (2023). Does using none-of-the-Above (NOTA) hurt students' confidence?. *Journal of Intelligence*, *11*, 157.
https://doi.org/10.3390/jintelligence11080157

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43*, 14–26. https://doi.org/10.3758/s13421-014-0452-8

Pachai, M. V., DiBattista, D., & Kim, J. A. (2015). A Systematic Assessment of "None of the Above" on Multiple Choice Tests in a First Year Psychology Classroom. *Canadian Journal for the Scholarship of Teaching and Learning*, *6*, 2. http://ir.lib.uwo.ca/cjsotl_rcacea/vol6/iss3/2

Papenberg, M., & Musch, J. (2017). Of small beauties and large beasts: The quality of distractors on multiple-choice tests is more important than their quantity. *Applied Measurement in Education*, *30*, 273–286.
https://doi.org/10.1080/08957347.2017.1353987

Rich, C. E., & Johanson, G. A. (1990). *An item-level analysis of "none of the above".* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Rodriguez, M. C. (1997). *The art and science of item-writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Sanderson, P. J. (2010). *Multiple-choice questions: A linguistic investigation of difficulty for first-language and second-language students*. Unpublished doctoral dissertation, University of South Africa, Pretoria.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distracters in multiple-choice questions: a descriptive analysis. *BMC Medical Education, 9*, 1–8. https://doi.org/10.1186/1472-6920-9-40

Trepte, S., Loy, L. S., Schmitt, J. B., & Otto, S. (2017). Hohenheimer Inventar zum Politikwissen (HIP). *Diagnostica, 63*, 206–218. https://doi.org/10.1026/0012-1924/a000180

Wesman, A. G., & Bennett, G. K. (1946). The use of "none of these" as an option in test construction. *Journal of Educational Psychology, 37,* 541–549.
https://doi.org/10.1037/h0056815

Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Journal of Scholarship of Teaching and Learning, 16,* 1–14. https://doi.org/10.1037/stl0000062

Yousfi, S. (2005). Mythen und Paradoxien der klassischen Testtheorie (II). *Diagnostica, 51*, 55–66. https://doi.org/10.1026/0012-1924.51.2.55