A peer reviewed, open-access electronic journal: ISSN 1531-7714

Ensuring Breadth and Depth of Knowledge on Multiple-Choice Examinations for Board Certification

Heath Kincaid, *American Board of Obstetrics and Gynecology*
Anthony Moreno-Sparks, *American Board of Obstetrics and Gynecology*
Pooja Shivraj, *American Board of Obstetrics and Gynecology*
Jill Holmes, *American Board of Obstetrics and Gynecology*
Amy Young, *American Board of Obstetrics and Gynecology*
George D. Wendel, Jr., *American Board of Obstetrics and Gynecology*

Abstract: Certification organizations aim to assess candidates on their breadth and depth of knowledge to determine eligibility for certification in their field of specialty. Assessments used for certification, when appropriately constructed, should use questions (or items) that assess the entirety of the field. However, comparing the plethora of the content of items to assess content coverage is a lengthy and time-consuming process. In an effort to become more aligned with the purpose of increasing content representativeness, organizations can implement a variety of Natural Language Processing (NLP) techniques with their items to ensure no one concept, medical condition, or scenario presents itself redundantly throughout each of its multiple-choice examinations. We provide an illustrative example from the American Board of Obstetrics and Gynecology (ABOG) of the NLP processes used to increase efficiencies and ensure content representativeness.

Keywords: Natural language processing, Board certification, Validity

Introduction

For many skilled professionals such as teachers and medical practitioners, certification accentuates the knowledge, judgement, and skills taken for minimum competency within a career field. However, certification organizations maintain the responsibility of defining minimum competency. Many certification organizations ensure a high standard of the minimally qualified candidate through practice analyses or Job Task Analyses (JTA), item writer trainings, standard settings, and upholding evidence-based psychometric standards. However, there lacks a substantial amount of literature on ensuring the breadth of the career field exists within the certification process, beyond tagging items to a content-relevant blueprint. While items may test differing concepts about a topic, a consistent presence of that singular topic could bias certification

decisions towards experts of such topic. For example, in the domain of obstetrics and gynecology (OB GYN) board certification in the United States, if too many items test abdominal hysterectomy, then certification decisions would bias proficiency in this particular procedure while missing other relevant content.

Psychometrically, assessing similar concepts about a topic can also violate local independence, an assumption of item independence to ensure that candidate solving algorithms cannot be copied from one item to another. A lack of local independence can lead to false assurances of quality in assessing person ability and item difficulty within psychometric modeling (Bond et al., 2021). Item pairs that violate the assumption of local independence are known as enemy items.

One option to ensure adequate content coverage and minimize enemy item pairs utilizes principles from data science. Specifically, Natural Language Processing (NLP) can be utilized to decipher the context of item similarity. NLP is defined as an area of research and application that explores the implementation of computational methodologies to understand and manipulate natural language text or speech to do useful things (Chowdhury, 2003). Typically, NLP analyses decipher sentiment and gather overall summary data from a collection of texts. More recently, NLP has been utilized in the space of text generation, notably with the rise of ChatGPT. However, we seek to adapt previous NLP methodologies in the analysis of source and target text, where in analysis of source and target text seeks to establish evidence of plagiarism, we seek to establish evidence of item similarity to ensure distinct construct representation within all written examinations.

We situate our work within content-related evidence of validity (American Educational Research Association et al., 2014). In particular, we aim to increase content reflective of knowledge, judgments and skills outlined within the test blueprint while decreasing overlap between common items. Current literature on NLP review systems follows a general procedure of removing erroneous formatting and common English text, normalizing English text by reducing gerunds and participles to a root form, and calculating a distance between each pairwise set of items within a form or bank. For example, current research typically performs a cosine similarity index of the stem text, the key text, and the concatenated stem and key text to flag item enemies for a SME review (Becker & Kao, 2022; Mao et al., 2021). Additionally, other research has proven successful in implementing computer-based algorithms like Latent Dirichlet Allocation, Latent Semantic Analysis, and Artificial Neural Networks to mimic SME judgement of item enemy identification with iterative improvement from SMEs (Weir, 2019; Peng, 2020). Each methodological approach possesses tradeoffs, yet integrated SMEs within the process to verify, improve, and serve as assistant for human intervention.

Operationally, NLP gives content editors/developers the tools to monitor the content of like items for similarity and replace redundant items with content that may not currently be present from the blueprint. In the current study, we utilized the American Board of Obstetrics and Gynecology (ABOG) written examination as an illustrative example to explore the application and implementation of NLP techniques in the enemy identification task. We aim to illustrate NLP implementation as a tool for assessment professionals to assist in achieving adequate construct representation in conjunction with blueprint mapping, creating additional assurances in the assessment of a breadth of content coverage within all ABOG multiple-choice examinations. We believe NLP methodologies along with SME reviews streamline and elevate the process of item-enemy identification by allowing for more targeted reviews of potential item conflicts, providing an additional layer of evaluation beyond blueprint classification, and solidifies a documented, objective procedure for form building processes.

We aimed to improve on previously used methodology at ABOG, which estimated similarity on the quantity of shared terms, and provide a holistic similarity algorithm based on preexisting NLP preprocessing

and analytical techniques. In the next section, we describe ABOG and their assessment system for initial board certification in obstetrics and gynecology in addition to the methods of the NLP analyses.

Method

ABOG provides initial and continuing board certification for Obstetrics and Gynecology (OB GYN) physicians. Initial certification is a two-phase process in which a candidate must receive a passing score on a computer-based, multiple-choice Qualifying Examination (QE), and an oral, face-to-face Certifying Examination (CE; ABOG, 2023). NLP analyses were performed on the QE, which comprises of one fixed form of 180 multiple-choice items and 50 unscored field test items. Each item is comprised of four plausible distractors with one single best answer.

Prerequisites

In creating an NLP algorithm to assess content similarity, we made certain operational-based decisions to best fit the style and content of the items. Specifically, we decided to forgo usage of semantic language models like BERT and utilize a normalized-descriptive approach due to initial trials having an ordinal similarity metric that did not align with the judgements of those with content similarity expertise in the domain of OB GYN. Additionally, we compiled words determined by SMEs that add little information to the overall content of an item to differentiate between unspecific content existing across the entire construct and content that differentiates distinct content within the construct. For example, the words "evaluation", "diagnosis", "woman", and "patient" add little information in the context of differentiating certification content within OB GYN. Additionally, those looking to implement an NLP algorithm can aim to maximize computational and operational efficiency but targeting specific subsets of items relevant to the form-building process. However, in cases of utilizing a linear-on-the-fly testing (LOFT) or computerized adaptive testing (CAT) algorithm, an algorithm must examine all possible item comparisons to have a full examination of all potential instances of local independence.

Additionally, NLP algorithms assessing content similarity must differentiate similarity within the domain of interest and succeed at identifying similarity in a manner that follows existing human logic. In order to assure algorithm quality, human domain experts should verify the quality of the algorithm output based upon the preconceived notions of a domain expert who has completed the content similarity task. For larger item banks, NLP algorithms utilize a significant amount of computational resources. In order to create an efficient algorithm that provides results in a timely manner, the algorithm must utilize best coding practices including the utilization of joins over filters, preventing the calculation of mathematically intensive formulae unless needed (e.g., similarity metrics), and the use of parallel processing to divide the computational resources across the machine.

Items Selection

The ABOG QE forms are built iteratively, with items used to assess a candidate's knowledge, judgment, and skills being selected first (scored items) for the fixed form, followed by new items that could be used as scored items in the future, pending satisfactory psychometric performance (field test items). In selecting scored items, ABOG prioritizes achieving optimal blueprint allocation, a Rasch item measure within 1.39 logits of each respective examination cut-score, and an item point-biserial correlation greater than or equal to 0.1. The items that fit these criteria are compiled into a form with their stem and distractor text and then analyzed after utilizing NLP preprocessing techniques.

Data Preprocessing

While all ABOG items follow a general style and format, some items have information that is not relevant to the content of the item. Specifically, items will have extra spaces, capitalizations, and other coding that is associated with delivering the item. This information is deleted and reformatted to ensure erroneous information cannot inflate or deflate a comparison metric. ABOG utilized words determined by SMEs that add little information to the overall content of an item and utilized the tm package in R to eliminate common construct-specific and English words with little informative meaning, as shown in Appendix A (Feinerer & Hornik, 2023; Feinerer et al., 2008; R Core Team, 2023). For example, words "and," "for," and "but" give little insight into the topic of statements in English. These words were deleted in the analysis of the item as a whole and the comparison of the stems. These words were not deleted in the analysis of the distractors as these statements are already incomplete sentences and are assumed to be comprised of only pertinent information. Next, all punctuation was removed, followed by lemmatization, where each inflected word form was returned to its basic form (Korenius et al., 2004), by replacing any inflected words contained within the lemma-word pairings in the `textstem` package (Rinker, 2018a; Rinker, 2018b).

Deriving Thresholds

When creating an NLP algorithm, a researcher must navigate the wide variety of the field to determine the approach that best fits in identifying item enemies. Specifically, a researcher can measure the contents of an item with a character-based approach, which is based on the contiguous chain of characters length which are present in both strings, a term-based approach, which relies upon the term frequencies, or a semantic-based approach, which analyzes the context of words within a given text. After the researcher measured the within-item content, the process of measuring between-item content begins. During the item enemy detection process, a variety of metrics exists to measure pairwise item sets. Specifically, the cosine similarity, Jaccard similarity, and Manhattan distance are all popular text similarity metrics that have unique capabilities to detect similarity, depending on the task and structure of the corpora of interest (Vijaymeena & Kavitha, 2016).

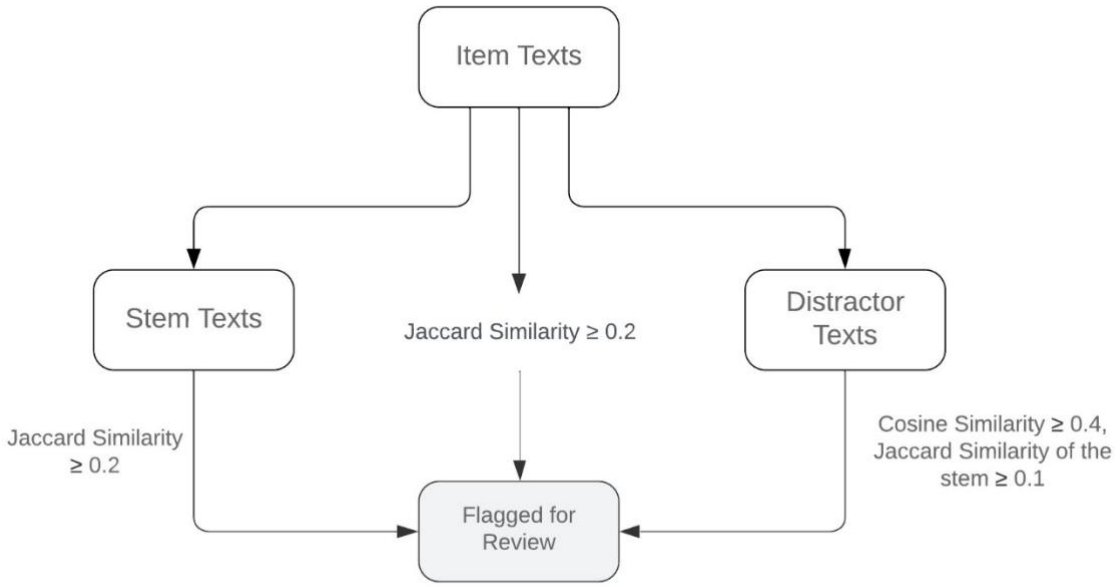
In our first iteration of an NLP algorithm, we utilized older items within the ABOG item bank to acquire minimum thresholds and evaluate the performance of comparison metrics. In this process, we utilized the 2022 Complex Family Planning form of 230 items and computed all possible pair-wise combinations of items. We utilize a term-based approach in which we calculated a Jaccard similarity index and a cosine similarity index, for analyses comparing the text of the stem and answer options, comparing the stem texts, and by comparing the concatenated text of the answer options. In previous literature, researchers suggest word repetition provides allows for linguistic cohesion across sentences, linking key ideas and allowing for greater understanding from the reader (He, 2014). Therefore, we chose to utilize the Jaccard similarity index in the analysis of the entire item texts and the stems of each question pair due to previous research associating word frequencies with linguistic cohesion. We postulated word frequencies to not be relevant in determining similarity in the entire item text or within the stem, which the cosine similarity metric considers. In contrast, we chose to utilize the cosine similarity index to account for the answer options due to each option being an independent clause. Given the mutual exclusiveness of each distractor, we hypothesized that the cosine similarity index could best represent pairwise distractor similarity due to the cosine similarity index deriving from frequency and word similarity, as opposed to the Jaccard similarity, which only utilizes the shared sample space of distinct words.

These metrics were then provided to the ABOG Assessment Development Team, who evaluated the comparison metrics of the stem and distractors in descending order, giving dichotomous labels of the comparison needing to be reviewed by an SME, or a dissimilar item comparison. In this trial analysis, we found that items started to become dissimilar once the Jaccard similarity of the stem approached 0.3. We also found a lack of a strong relationship between a higher cosine similarity score between item distractors

and content similarity. However, we found the all-item text similarity could indicate content interaction between the stem and distractors. Following this anecdotal analysis, we decided to include all three analyses holistically to review content similarity, with an emphasis on reviewing item comparisons with high Jaccard similarity metrics for the all-item and stem texts.

Upon completion of the trial analysis, comparisons flags were postulated based off the trichotomous labeling from the ABOG Assessment Development Team of "Similar", "Not similar", or "Needs SME reviewal". As shown in Figure 1, we flagged comparisons if the Jaccard similarity of their stem was greater than or equal to 0.2, as comparisons appeared to be dissimilar below 0.3 and needed SME reviewal between 0.3 and 0.4, and similar comparisons above 0.4. We postulated a threshold of 0.3 for needing reviewal by a content editor or an SME, however we decided to lower the threshold to 0.2 to allow for margin of error. An item comparison was also flagged if the cosine similarity of the distractors was greater than or equal to 0.4 and the Jaccard similarity of stem was greater than or equal to 0.1. This criterion was motivated by many comparisons with high cosine similarity indices not being similar items as the verbiage of the stem assessed different concepts. However, if a comparison had a slight degree of similarity in the stem along with a great degree of similarity in the distractors, we valued a qualitative review from a content expert. The final metric threshold of the stem and distractor text combined was set at a Jaccard similarity index of 0.2. This was motivated to include comparisons that were just under 0.2 threshold for being flagged for similar stems but could have some interaction between the stem and the distractors that could make the overall item content similar.

Figure 1. Item Similarity Criterion to be Flagged for Manual Review

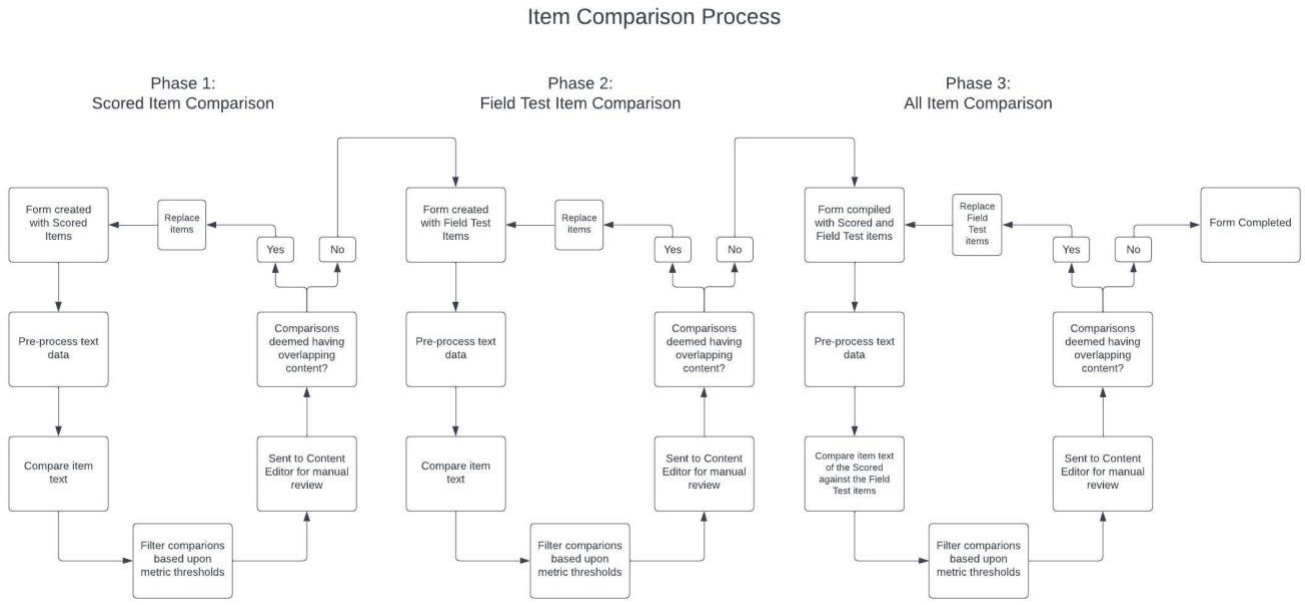


Analysis

With the thresholds set above, the item comparison process was utilized to build each form for all ABOG QEs in 2023. The analysis was sent to the Assessment Development Team to review. Any item pairing deemed to be excessively similar would have the item with less desirable psychometric statistics deleted from the form and would then have a different item added to the form. Once all items have been replaced, the form will be re-analyzed for similarity, subsetting all comparisons to the ones in which a new item is included. As shown in Figure 2, this process is completed iteratively until there are no more items

deemed to be similar by the Assessment Development Team, then this process will be completed for the field test items and finalization will occur by comparing the field test items to the scored items.

Figure 2. Process Chart defining the steps in ensuring distinct content on ABOG Written Examinations



Results

In the first year of operationally using NLP analyses to assist in identifying similar content, we analyzed 75 different iterations of item sets across all six content areas. Overall, we algorithmically assessed 522,247 pairwise combinations of items, stemming from 2,426 distinct items to ensure a breadth of content coverage existing within the ABOG QEs. Upon NLP analyses, the ABOG Assessment Development team reviewed 1,682 comparisons to establish the presence the content similarity. From the manually reviewed comparisons, we instantiated 437 distinct item substitutions. Each substitution resolved at least one enemy conflict; however not all enemy decisions were recorded. Overall, our analyses handled 99.7% of all item comparisons, identifying item enemies at a rate of 16.3%, without any SME feedback (Table 1).

Table 1. Reduction in Manual Comparison When Using Text Similarity Analyses

Exam	Items on Form	Items Evaluated	Algorithmic Comparisons	Manual Comparisons	% Analyzed by Algorithm Only	% Items Removed by Algorithm
CFP	230	244	12,161	143	98.8%	5.7%
FPMRS	280	342	55,223	437	99.2%	18.1%
GYN ONC	280	362	55,469	239	99.6%	22.7%
MFM	280	335	51,017	149	99.7%	16.4%
REI	280	337	54,766	122	99.8%	16.9%
Specialty	680	806	293,611	592	99.8%	15.6%

While SMEs would not typically review these items pairwise in a way that this analysis does, finding similar items without the assistance of an NLP algorithm is quite challenging and non-robust, which leads a high chance that items with similar content and scenarios are placed with the same exam. The SME workload needed to identify these items can become burdensome, leading to less credibility that the assumption of a breadth of knowledge is needed to pass an exam and could skew the perception of the definition of a minimally qualified candidate in a respective certification program, lowering the standards for OB GYN care.

Conclusion

The role of a certification organization is to assess candidates on a wide variety of knowledge, judgement, and skills to verify a skillset of interest. In staying aligned with these ideals, certification organizations must be cognizant of the overrepresentation of content within examinations while ensuring sufficient content coverage from the required skillset. NLP has the opportunity to mitigate two main issues in the exam development process: item dependency and time to build assessments.

Item Dependency

Utilization of the current NLP methodology mitigates issues of item dependency that are typically found after test administration by providing a concise procedure that standardizes text, and quantifies similarity. NLP creates a manageable workload for SMEs to efficiently rid of item dependencies with more accuracy and precision, increasing psychometric defensibility. Despite the differences used between NLP text-similarity systems, a certification organization must utilize an SME to determine what constitutes an enemy item pairing due to the distinct semantic nature of each topic. However, NLP text-similarity systems can provide a better and more efficient review without having to factor in the time of an SME and common errors associated with human review.

Time to Build Assessments

While organizations may choose to implement manual review after NLP analyses, the number of items to compare can be significantly reduced via standardizing information in the item text and comparing items in an iterative, computational manner. In the current study, we found NLP analyses provided targeted feedback to SMEs at a 99.7% reduction. This standardized process then shifts more time for content editors and SMEs to make judgements about item quality and find any additional item with similarity that cannot be assessed via similar wording, improving the quality and defensibility of certification decisions.

Future Research and Implications

The steps outlined here can be improved in the future to better identify similar content. Specifically, the lemmatization lexicon can be improved to include context-specific terminology, as well an additional analysis to determine if common words appear together in a sequence, which could be more suggestive of an underlying topic as opposed to only analyzing if the items share similar words. Additionally, the application of a Boolean variable for an identical key between variables could provide additional context of item similarity, as consistent keying across items could limit construct representation across a domain.

While the aforementioned analysis presents a specific use-case in the domain of obstetrics and gynecology certification, testing professionals can incorporate NLP analyses to optimize item banking practices, with certain caveats. Specifically, each domain has a very specific lexicon with differing underlying semantic relationships, requiring operational research to explore the relationship between similarity metrics and the item enemy label derived by SMEs. For example, each subject possesses domain-specific stop words beyond common English phrasing, that must be identified in removed in analysis to give a more accurate

assessment on similarities between item themes. While the analysis here presents a bag-of-words approach in deriving item similarity, other methodological approaches may perform more desirable depending on the domain. Specifically, Large Language Models (LLMs) can derive a semantic meaning approach in item similarity, in which semantic embeddings generated from a LLM are compared via a cosine similarity metric. However, this approach still does not yield a direct identification of item enemies, and a manual review item review is needed to ascertain the definition of an item enemy.

In the measurement field, item enemy detection systems are vital in the implementation of Linear-on-the-fly testing (LOFT), computer adaptive testing (CAT), and item response theory (IRT) models. Automated forms of form building like LOFT or CAT rely on specific test specifications dependent upon blueprint allocations, and, in the case of CAT, psychometric measures to ensure comparable difficulty across the form with items of varying content presented. In order to execute LOFT, CAT, and comply with the local independence assumption within an IRT model, certification organizations must identify item enemies in an efficient fashion due to an overwhelming number of potential comparisons that typically exist within an item bank, which leads to most look towards a systemic and efficient process like an NLP text-similarity system.

NLP can be extended for additional testing developing uses including other types of examinations. For example, oral examinations in which test takers respond to prompts can also be evaluated for similarity across prompts. Additionally, any examination where examinees submit documentation for testing can also be evaluated for similarity. For example, an aspect of the ABOG oral examination includes the submission of case list provided by OB GYN physicians. These case lists can be evaluated for similarity to ensure that examinees are submitting unique case lists and that examinees are each meeting the requirements for examination, thus providing evidence of fairness in assessment. Developing tests involves a rigorous process of identifying items, comparing items for similarity, and selecting the appropriate items for the forms. If multiple items are found to assess the same content on an examination, then assessment developers should consider additional items to address gaps in the test blueprint. While manually comparing items is an option for assessing similarity, NLP assists the evaluation content representativeness of items on exams while decreasing the time needed during manual comparison. As NLP continues to create additional tools to optimize and enhance language analysis, certification organizations must commit to evolving with the ever-changing best practices, ensuring a wide variety of knowledge is needed to be board-certified, and making certification decisions an accurate representation of a breadth of knowledge, judgement, and skills.

Received: 6/3/2024. **Accepted:** 12/1/2025. **Published:** 12/9/2025.

Citation: Kincaid, H., Moreno-Sparks, A., Shivraj, P., Holmes, J., Young, A., & Wendel Jr., G. D. (2025). Ensuring breadth and depth of knowledge on multiple-choice examinations for board certification. *Practical Assessment, Research, & Evaluation*, 30(1)(12). Available online: https://doi.org/10.7275/pare.2117

Corresponding Author: Heath Kincaid, American Board of Obstetrics and Gynecology.
Email: hkincaid@abog.org

References

American Board of Obstetrics and Gynecology (ABOG) Specialty Certification Bulletin. Accessed September 30, 2023. https://www.abog.org/specialty-certification/bulletins

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Becker, K. A., & Kao, S. C. (2022). Identifying enemy item pairs using natural language processing. *Journal of Applied Testing Technology*, 41-52.

Bond, T.G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental measurement in the human Sciences*. Routledge.

Chowdhury, G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37. pp. 51-89.

Feinerer, I. & Hornik, K. (2023). `tm`: Text Mining Package. R package version 0.7-11, https://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54. doi: https://doi.org/10.18637/jss.v025.i05

He, Q. (2014). Implications of lexical repetition patterns for language teaching. *International Journal of Linguistics, 6*(4), 46–58. https://doi.org/10.5296/ijl.v6i4.6115

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004, November). Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 625-633).

Mao, X., Zhang, Q., & Clem, A. (2021). An Exploration of an Integrated Approach for Enemy Item Identification. *International Journal of Intelligent Technologies and Applied Statistics*, 14(2), 123-134. https://doi.org/10.6148/IJITAS.202106_14(2).0004

Peng, F. (2020). *Automatic enemy item detection using natural language processing* (Doctoral dissertation). University of Illinois at Chicago.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computer.

Rinker, T. W. (2018a). lexicon: Lexicon Data version 1.2.1. http://github.com/trinker/lexicon

Rinker, T. W. (2018b). `textstem`: Tools for stemming and lemmatizing text version 0.1.4. Buffalo, New York. http://github.com/trinker/textstem

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal, 3*(2), 19-28.

Weir, J. B., II. (2019). *Enemy item detection using data mining methods* (Doctoral dissertation). University of North Carolina at Greensboro.

Appendix A

Stopwords Utilized in ABOG Content Similarity Analyses

Stopwords				
a	copays	host	operation	stopped
abandon	correctly	hot	opinion	store
ability	cotton	hour	optimally	stores
able	could	hours	option	straw
about	couldn't	house	options	strong
above	count	how	or	strongly
absolute	create	how's	other	studied
accept	criminal	human	others	study
accepted	cross	I	otherwise	subject
accompany	cup	I'd	ought	success
acquire	daily	I'll	our	such
acquired	dark	I'm	ours	sudden
across	date	I've	ourselves	suddenly
action	day	ia	out	suggest
active	days	icu	outline	suggests
activity	decide	ideally	outside	summary
add	decided	if	outweigh	sure
added	decrease	iii	over	switch
admit	decreases	importance	overall	syrup
admitted	deep	important	own	t
advise	deg	improve	oxygen	take
advised	degree	improved	pack	taken
after	deploy	improves	pair	taking
again	despite	in	part	tan
against	develop	inch	parts	taste
age	device	inches	past	team
ages	devices	include	patient	teenager
ago	did	included	pay	tell

agree	didn't	includes	per	tells
agrees	die	increase	perform	ten
aid	died	increased	performed	test
air	discuss	indeed	phase	testing
all	discussion	indication	phone	tests
allow	display	influence	physician	than
almost	divide	informative	piece	that
alone	division	initial	place	that's
along	dm	initially	plain	the
also	do	inquire	platelet	their
am	doctor	inside	play	theirs
among	document	instead	plays	them
amount	does	instruction	poorly	themselves
an	doesn't	intellectual	possible	then
and	doing	intend	power	there
another	don't	intention	practice	there's
answer	done	into	prep	these
any	down	is	prepare	they
anyone	draw	isn't	prescriber	they'd
anytime	due	it	present	they'll
appear	dump	it's	pressure	they're
appendix	during	item	previous	they've
apply	each	its	previously	thing
appointment	earliest	itself	primarily	think
approach	early	just	prior	third
appropriate	easily	justify	problem	thirty
appropriately	eat	k	proceed	this
are	effect	kg	process	those
area	effective	kind	produce	though
areas	effectiveness	kit	produced	thought
aren't	effects	know	progress	three
around	eight	knowledge	prolong	through

arrive	eighth	known	prompt	throughout
arrives	either	l	propose	time
as	enable	label	provide	timed
ask	end	last	provided	times
asked	ends	lasted	pt	to
asking	english	lasting	pull	today
asks	enough	late	push	together
associate	enter	later	put	told
associated	equal	lb	question	too
assume	essentially	lead	questions	took
at	estimate	leading	quiet	top
attach	estimates	leaf	quietly	topic
attempt	even	least	quite	total
attempting	ever	leave	raise	toward
aunt	every	less	rarely	town
author	exam	let's	rate	travel
authority	examine	letter	rather	treat
available	examined	level	rdquo	tried
average	exams	levels	reach	true
avoid	exceed	light	reached	try
avoided	exceeds	like	read	trying
await	except	likely	reads	turn
away	exhibit	limit	ready	twelve
back	exist	limiting	real	twenty
bad	expect	list	rearrangement	twice
bag	expected	listed	reason	two
banana	explain	liter	reasons	typical
base	explains	little	receive	typically
based	explore	locate	received	unable
basic	expose	located	receiving	under
be	exposes	long	recommendation	undergo
bear	extra	look	recommendations	undergone

became	extremely	lose	record	unknown
because	eye	lost	red	unless
become	f	lot	reduce	unlike
becomes	fact	low	reduces	unsuccessful
becoming	failure	lower	refer	unsure
bed	failures	lowest	refers	until
been	false	m2	reflect	up
before	family	made	regular	upcoming
began	fan	main	release	update
begin	far	make	released	upon
beginning	fast	maker	relies	urge
begins	fat	male	relieve	us
begun	favor	man	rely	use
behind	feature	manage	remain	used
being	features	managed	repeat	useful
below	feel	management	request	user
best	feels	manner	requested	users
beta	female	many	resolve	uses
better	few	marathon	resolved	using
between	fifth	mark	respect	usual
beyond	figure	markedly	respond	usually
blanket	file	marriage	result	value
blood	fill	may	results	values
board	fills	me	return	very
body	financial	mean	returns	visit
both	find	means	reveal	voice
bpm	finding	meet	reveals	w
brief	fine	member	ride	waist
bright	fingernail	men	right	wait
bring	first	met	risk	waive
brought	fit	method	risks	want
build	fits	methods	room	wants

burn	five	mg	root	warn
but	flight	might	routine	was
by	fold	mild	rsquo	wasn't
c	follow	military	rule	watchful
call	followed	minus	runner	way
called	following	miss	safe	we
calls	for	missed	safest	we'd
camp	form	miu	same	we'll
can	forty	mix	saturation	we're
can't	found	mixed	say	we've
cannot	four	ml	says	weak
capable	fourteen	mm	schedule	weakest
care	free	mmhg	scheduled	weakly
careful	frequent	mmol	school	weapon
carefully	friend	model	seat	web
carried	from	mom	see	week
carries	full	month	seek	week's
carry	function	more	seeking	weeks
case	further	most	seeks	well
cast	g	mostly	seen	went
catch	gather	motion	sees	were
cause	general	move	select	weren't
causing	generally	moved	self	what
center	genuine	much	send	what's
chain	gestation	must	sent	when
chance	get	mustn't	serve	when's
change	gift	my	set	where
changes	girl	myself	sex	where's
chart	give	natural	shan't	whether
check	given	near	share	which
checked	gives	nearly	she	while
checklist	go	need	she'd	whirlpool

chew	good	needed	she'll	white
choice	great	needing	she's	whiteboard
choices	greatest	needs	shellfish	who
choose	gum	never	shoot	who's
climb	had	new	should	whole
clinician	hadn't	next	shouldn't	whom
close	half	no	show	whose
closely	handle	noise	showing	why
cm	handwriting	none	shown	why's
cohort	happen	nor	shows	wife
colleague	happens	normally	side	will
colleague's	hart	not	sided	wind
collision	has	note	similar	with
come	hasn't	noted	simultaneously	within
comes	have	now	since	without
comet	haven't	numb	sit	witness
commercial	having	number	six	witnessed
common	hcg	observe	size	woman
commonly	he	obviously	sized	women
community	he'd	occur	slightly	won't
company	he'll	occurred	slow	work
compete	he's	occurs	slowly	works
competed	hear	odd	so	world
complain	help	of	soap	worse
complaint	helpful	off	social	worsen
complete	her	offer	some	would
completely	here	offered	something	wouldn't
comprehensive	here's	offers	soon	write
computerize	hers	office	space	year
confidence	herself	officer	speed	years
confirm	high	often	staff	you
connect	him	old	stair	you'd

consider	himself	on	start	you'll
considered	his	once	started	you're
consumer	history	one	statement	you've
contact	hit	only	stay	your
contain	hits	onset	step	yours
contains	hoffman	onto	still	yourself
content	hold	open	stop	yourselves
control	home			